EXECUTIVE SUMMARY
of Report PM 98-01 Entitled
"A Computer Program to Identify Outliers in the Pesticide Use Report Database"

California Environmental Protection Agency
Department of Pesticide Regulation
Environmental Monitoring and Pest Management Branch

PURPOSE

The Department of Pesticide Regulation (DPR) maintains a database of California pesticide use information as reported by all agricultural users and by structural pest control businesses. Considering the vast amount of data, the percentage of errors is small. However, even a small number of errors can significantly affect the accuracy of analysis if these errors represent overstatements of pesticide use by several orders of magnitude. (For example, an application of 128 pounds of active ingredient is made but as a result of a decimal point shift by the user in hand-entering the data, the use report reflects 128,000 pounds of active ingredient--127,872 pounds too much. Note: under reporting of applications is of less concern. If the decimal point shift were in the other direction and .128 pounds were reported, the reported pounds would only be 127.9 pounds too little. It would take 1000 such incidences of under reporting to equal the impacts of the one example of over reporting.) Since large errors reduce the confidence of any analysis that uses the database, DPR developed a means to minimize the number and magnitude of the errors in the pesticide use reporting (PUR) database.

BACKGROUND

In 1990, California enacted legislation requiring that all agricultural pesticide use be reported to DPR. Agricultural use includes applications to crops, parks, cemeteries, golf courses, and rights of way, such as roadsides and railroads. In addition, all applications made by residential and structural pest control businesses must be reported. Typically, users submit 2.5 million records each year, and each PUR record contains 15 pieces of information on every pesticide application by commercial pest control operators and 30 pieces on every pesticide application by growers in California. Many people would like to use the PUR to analyze pesticide use for various purposes. This interest is increasing with the 1996 passage of the federal Food Quality Protection Act, which requires the U. S. Environmental Protection Agency to characterize overall pesticide risk, taking into account how pesticides are used. In addition, DPR uses pesticide use data to improve estimates of dietary risk, to locate sites for monitoring pesticides in the environment, to ensure compliance with clean air plans and ground water regulations, to assist county agricultural commissioners (CAC) in protecting endangered species, and to help identify reduced-risk pest management alternatives for specific crops grown in different regions of the state.

The best way to ensure high quality PUR data is to check use report data accuracy before data are entered into the database and then make sure the data are entered accurately. That process is facilitated when growers and pest control operators use the California Electronic Data Transfer System to submit PUR data directly to counties in electronic form. Otherwise, pesticide users submit paper reports to the CAC whose staff then enter the data in electronic format. DPR's Information Systems Branch (ISB) has developed a program to help CAC staff screen out errors in reported locations of applications, commodity treated, acres planted and treated, identification of the operator, and identification of the pesticide applied, among others. CACs also use this program to identify illegal uses. The PUR data is also screened by ISB in Sacramento to make sure the commodity treated is a legal use of the reported pesticide and, more recently, to identify many errors caused by reporting of extremely large pesticide use rates. However, at this time, the ISB methodology does not identify all extremely large reported application rates that are possible errors. Thus, DPR needs to refine the program to include criteria that can be used to screen extremely large application rates that may be errors. These same criteria could also be used to flag historical PUR data that was entered before the current ISB screen for extremely large pesticide use rates was developed. When these historical data are used in an analysis, possible errors can be included or not, depending on the type of analysis, the pesticide(s) involved, and knowledge of the analyst.

Theoretically, extremely large values that are errors could be identified by comparing maximum label rates with application rates in the PUR. However, maximum label application rates are not currently available in an easily accessible database. ISB is evaluating several existing systems to include this information in future enhancements to the county/state PUR validation process.

STUDY METHODS

The Environmental Monitoring and Pest Management Branch analyzed pesticide use records from 1991-1995 for possible outliers. Only extremely large values were considered possible outliers because these values can greatly distort total pesticide use figures.

Five criteria were evaluated to identify possible outliers: four to identify errors in pesticide use rates and one to identify errors in acreage treated.

Use rates

Four criteria were used to flag records as possible errors in PUR use rates.

(1) Criterion 1 flagged use reports that exceeded specified pounds of pesticide applied. A lower threshold value was set for nonfumigant pesticides, a higher value for fumigants which are applied at much higher rates than nonfumigants.

(2) Criterion 2 flagged use reports that exceeded the median value of all similar applications by a specified amount.

(3) Criterion 3 flagged use reports that exceeded the median value plus a measure of variation of all similar applications by a specified amount.

(4) Criterion 4 flagged use reports that exceeded threshold values generated by a neural network.

A neural network is a mathematical function that calculates a set of output values from a set of input values. To do this, the function has a large number of parameters that are set so that the network will give the correct outputs for every possible set of inputs. The parameters are set by “training” the neural network, that is, by presenting the network with a set of data consisting of many sets of input and corresponding output values. The neural network program then adjusts the parameters so that it produces the correct output values for each input set.

Area treated

One criterion was used to flag records as possible errors in PUR "area treated."

(5) Criterion 5 flagged use reports in which reported acres exceeded 700.

Analysis

The outlier criteria can make mistakes by either flagging records that are not really outliers (type I errors) or by overlooking outliers (type II errors). To be conservative, that is, to minimize the exclusion of valid records from an analysis, the goal is to minimize type I errors. Each of the four use rate criteria was evaluated to determine the situations in which it (1) worked well**,** (2) made type I errors, and (3) made type II errors. In addition, each criteria and one combination of criteria were applied to each PUR for 1991-1995 to determine the number and percentage of records that were outliers. Selected criteria were also used to determine, for the 1995 PUR, the percentage of outliers by county and by active ingredient, and to determine the percentage change in total pesticide use by county and active ingredient after deleting outlier records.

RESULTS

None of the criteria worked in every situation, but in general criterion 4 (neural networks) was best at identifying outliers over the broadest range of situations. Criteria 1 and 2 failed to identify many records that were obviously outliers. Criterion 3 worked well for normal (bell-shaped) distributions of reported use, which are rare in pesticide use, but flagged too many valid records with non-normal distributions to be used uncritically. Criterion 5 (more than 700 acres) identified the fewest outliers which was to be expected because this criterion is used to screen data before they are entered into the PUR.
However, each criterion can find some outliers that the others cannot in specific situations. Combinations of criteria, such as a specific 1, 2, and 4 combination,

appeared to give the best results when analyzing a variety of pesticide use situations.

Percent of records flagged as outliers for all of California for 1991-1995. Using criteria 1, 2, and 4, the statewide outliers ranged from 0.56 to 0.83 percent of the total number of pesticide records checked and tended to decrease between 1991 and 1995.

Percent of records flagged as outliers by county for 1995. Using criteria 1, 2, and 4, the percent outliers by county ranged from 0 to 13.5 percent for individual years, and from 0 to 4.3 percent averaged over 1991-1995. The median percent outliers by county averaged over 1991-1995 was 0.56 percent. Urban counties tended to have a higher percentage of outliers than agricultural counties. Otherwise, no counties had consistently more or fewer outliers. However, as noted below even one extremely large outlier can greatly distort analyses based on total weight of pesticides applied.

Percent of records flagged as outliers by individual active ingredient. Using criteria 1, 2, and 4, the percent outliers by active ingredient for the top 50 active ingredients averaged over 1991-1995 ranges from 0 to 92 percent for individual years, and from 3.3 to 14.6 percent averaged over the five years. Many of these pesticides were somewhat special, unusual, or used in non-agricultural sites. They included alcohols, sex pheromones, bleach, garlic, soap, sawdust, insect and plant hormones, biologicals, and fumigants. In the case of fumigants, many valid records may have been identified as outliers because the criteria values were set too low.

Change in pounds of active ingredient for each county. In each of six counties total pounds of active ingredient used in 1995 increased by more than five percent when outliers identified by criteria 1, 2, and 4 were added. In each of two counties total pounds increased by more than 10 percent. However, counties with relatively high percentage changes in pounds of active ingredients did not correlate highly with counties with relatively high percentages of outlier records. This suggests that there are probably just a few very extreme outliers.

Change in pounds of active ingredient for each active ingredient. The effect of identifying outliers is most dramatic when calculating the total number of pounds of individual active ingredients. When outliers were added using criteria 1, 2, and 4, the total number of pounds of active ingredient for 1995 increased by more than 1000 percent for eight active ingredients and by a median of 37 percent for an additional 42 active ingredients. The largest change, 6900 percent, occurred when outliers for *Agrobacterium radiobacter*, a biological pesticide, were included. Including outliers identified by using criterion 1 increased reported use of carbaryl in the state from 0.8 million pounds to 1.5 million pounds. This change is due to a single extreme outlier value. This record was confirmed to be an error and corrected.

CONCLUSIONS AND RECOMMENDATIONS

Each of the four pesticide use rate criteria evaluated can be used to identify outliers depending on the particular use rate characteristics of a pesticide in the PUR. In general, criterion 4 and the combination of criteria 1, 2, and 4 were the most accurate criteria. The percent of outliers expressed as number of use reports was usually less than one percent, both statewide from 1991-1995 and in individual counties for 1995 (the only year analyzed by county).

The percent of outliers expressed as total pounds of active ingredients ranged from 5-10 percent in the top eight counties. However, the impact of outlier analysis was greatest with total pounds of active ingredient reported by individual active ingredient. Total reported use could be overstated by more than 20 percent for many individual active ingredients and by more than 1000 percent in a few cases, demonstrating the critical importance of outlier analysis of the PUR.

If the pesticide use being analyzed is characterized by a given distribution of pesticide use (examples: bell-shaped or bimodal distributions), then the criterion that best fits that distribution can be used. However, if the distribution of pesticide use being analyzed cannot be characterized or is characterized by a variety of distributions, it may be advantageous to use a combination of criteria, such as criteria 1, 2, and 4. If a quick analysis is necessary, only the most extreme, and thus the most certain, outliers should be excluded from the analysis. If a more detailed analysis is necessary, outliers identified by less extreme criteria could be examined to determine with more confidence whether or not they are truly errors.

Improvements can be made in the outlier procedures. For example, criterion 4 could be improved by using a larger training set and by testing different training procedures. Criterion 1 could be improved by setting higher criteria values for fumigants. Also, these criteria were only used to screen records with rates of use, such as pounds per acre, which require reports of the number of units treated (e.g., acres). But many records in the PUR have no information about the units treated. Other outlier criteria need to be developed for pesticide records with no unit data.

The presence of even one outlier can seriously affect a use analysis, which demonstrates the importance of identifying outliers in the PUR.

These new outlier criteria will be used to refine the program ISB uses to identify extremely large application rates that are errors in future entries in the PUR. In addition, DPR will use the criteria to flag possible errors in the PUR from 1990-1995 for use in future analyses of these data.

Doug Okumura
Branch Chief

A Computer Program to Identify Outliers in the Pesticide Use Report Database

By

Larry Wilhoit

April 1998

STATE OF CALIFORNIA
Environmental Protection Agency
Department of Pesticide Regulation
Environmental Monitoring and Pest Management Branch
Pest Management Analysis and Planning Program
Sacramento, California 95814-5624

PM 98-01

Abstract

The Department of Pesticide Regulation's (DPR) Pesticide Use Report Database (PUR) is an invaluable resource of information on the patterns of pesticide use in California. Many people both within and outside DPR use this database. However, as with any database in which much of the information is entered by hand, there are bound to be mistakes. Large errors reduce the confidence of any analysis that uses the database. Thus it is critical that the number and magnitude of errors be minimized.

The best procedure would be to prevent errors in the first place, but for the existing data in the PUR from past years this is no longer a possibility. The only option for past years is to mark records as possible errors. When these data are used in an analysis, one can then decide which possible errors to include or not based on whatever available knowledge one has.

This memo describes four methods for determining which pesticide use rates in the PUR are possible errors and one method for determining if the acres treated were in error. These procedures identify possible errors by comparing each use rate with an estimate of a reasonable rate for that type of use. If any rate is unusually high it is marked as an outlier, and thus a possible error, in the database. Four different types of procedures (or criteria) were used to identify outliers. One criterion compared each rate to a fixed maximum pounds of active ingredient per acre, a second criterion compared each rate to the median pounds of pesticide product per unit area treated for similar uses, a third criterion compared each rate to the median value plus a measure of variation in use, and a fourth criterion used a neural network procedure to identify outliers. A neural network is a special kind of function that can be used to estimate values that are determined by a complex interaction of many different factors. A final, fifth criterion, identified outliers not in rate of use but in number of acres treated. Records were marked by this criterion if the number of acres treated was greater than 700. These procedures were programmed in Oracle so that individual records in the Oracle PUR database could be marked.

These procedures were carried out on all the data in the PUR for the years 1991 through 1995. The results of this procedure were examined to determine how well each criteria worked in identifying outliers, to calculate the number of outliers found in the PUR, and to determine the effect the presence of these outliers would have on different kinds of pesticide use analyses.

Based on an examination of a sample of different pesticides and crops, it was concluded that at least some of these procedures were successful in correctly identifying outliers in the PUR. Each criteria had some shortcomings in certain situations, but for most situations the neural network criterion worked very well in identifying outliers. However, some of the other criteria, especially the first criterion, could be used to find some kinds of outliers that the neural network missed.

In general the percentage of all records in the PUR that were clearly outliers was somewhat less than 1%. However, many of the outliers found were extremely large;

some were millions of times the normal rates for that particular pesticide/crop situation. The effect of these extreme values was quite dramatic, especially in totaling the pounds of particular active ingredients. The total use of over a half a dozen active ingredients was changed by over 1000% by the presence of a few extreme records and the total use of many more active ingredients was affected by over 10%. These results demonstrate the seriousness of the outlier problem in the PUR.

Although these criteria were developed to identify outliers in the past years of the PUR, they could also be used to screen rate values as they are entered into the database by the counties. This would greatly reduce the number of extreme errors in future years of the PUR.

Table of Contents

List of Tables

List of Figures

Introduction

The Department of Pesticide Regulation's (DPR) Pesticide Use Report (PUR) is probably the largest and most complete database in the world on pesticide use for a major geographical region. Each record in the database contains a wealth of information on each and every pesticide application by growers and commercial pesticide control operators in California since 1990. Each year's data contains close to two million records. Many people would like to use this database for analysis of pesticide use for many different purposes. Many other states and countries are looking to this database as a model of how they might implement similar databases in their areas. This interest is only increasing with the passage of the Food Quality Protection Act, which requires U.S. EPA to make judgments on overall pesticide risk taking into account pesticide use.

Because of the importance of the PUR, it is critical that the database be as accurate and complete as possible. Many analyses require, in particular, information on pounds of pesticide product or active ingredient used and acres treated. A common need, for example, is to find the pounds of some active ingredient used on some crop in some region. If only one record has a large erroneous value, the total sum could deviate significantly from the true value, thus seriously effecting the analysis.

Whether or not a reported value for the pounds of product is an error can only be determined by checking the source of the value, which can realistically only be done at the time of data entry. It is impractical or impossible to find the original source of information for the records in the past years of the PUR. However, the most extreme (and detrimental) errors can probably be identified as extreme outliers in the distributions of values.

This memo describes several techniques for determining whether particular values of pounds of product or acres are outliers. I have written a computer program (attached to this memo as appendix III) to check the pounds of products used and acres treated for each record in the PUR. The program places a flag in one or more of eleven fields within the database if the values are determined to be outliers based on several different criteria. No records are removed from the database, but only flagged as outliers and thus potentially in error. Thus it is up to the user of the database to determine whether or not to include records flagged as outliers in their analysis. The program was run for every year of the PUR (1991 through 1995).

This memo also examines the frequency distribution of use rates of a few pesticides to determine whether the records flagged by the program are apparent outliers and whether the program fails to flag other records that are apparent outliers. The different criteria are compared to determine which appear to be most accurate in different situations.

Finally, the memo presents the percentages of the outliers for different counties and active ingredients and the effect of removing flagged records on the total pounds of pesticides used for different counties and active ingredients. These results might suggest where errors are most likely to arise.

Outlier criteria description

There are many possible methods for determining if a value is an outlier. If we knew the maximum label rates for particular uses, then rates in the PUR could be compared to these maximum rates, but unfortunately this information is not available in the PUR or in the Pesticide Label Database. The only other way to identify outliers involves looking at the distribution of the actual use rates. If the values are normally distributed then there are a number of statistical procedures for identifying outliers. If the values have an unknown or nonstandard distribution, then there exist no standard statistical procedures for identifying outliers. Nevertheless, people can usually look at a distribution and say with different degrees of confidence whether some value is an outlier. This suggests there should be some kind procedure that can be developed to make similar judgments. This memo will look at four different procedures or criteria for identifying outliers in the rates of use of pesticides. A fifth criterion looks at the total number of acres treated.

Only extremely large values are flagged, not extremely small ones, because only large values will distort sums. For each criteria, if a value is larger than some value (which will be called that criterion's "limit value"), then it will be flagged as an outlier What value to use for each limit value is somewhat arbitrary. Limit values were chosen for most criteria to be as close as possible to values that were considered to be outliers by a group of scientists in a survey described below in section on the neural network criteria. Since the limit values are somewhat arbitrary, each criteria had two or more limit values to flag different levels of outlier extremes.

The first four criteria used to identify outliers evaluate the pounds of pesticides applied per unit treated. The unit treated could be a measure of area (acres or square feet), volume (cubic feet), weight (pounds or tons), or some other unit such as number of tractors, trees, bins, etc. The first criterion examines only records with units treated in acres, but the other three criteria examine uses on all units treated. Also, the first criterion uses pounds of active ingredient whereas the other criteria use pounds of pesticide product.

The four criteria are briefly described here but a more complete explanation of each criterion is given in Appendix I.

<u>Criterion 1: Pounds per acre of active ingredient is larger than 200 or 400 (non-fumigants), or 1000 or 2000 (fumigants).</u>

Records were flagged in the PUR by criterion 1 if the pounds per acre of a non-fumigant active ingredient were greater than 200 or if the pounds per acre of a fumigant active ingredient were greater than 1000 (criterion 1a). Another field was flagged if the pounds per acre of a non-fumigant were greater than 400 or if the pounds per acre of a fumigant active ingredient were greater than 2000 (criterion 1b). These limit values were chosen based on what is known about typical rates of use for most pesticides.

Criterion 2: Pounds per unit treated of a product is larger than 25 or 50 times the median.

Records were flagged by criterion 2 if the pounds of pesticide product per unit treated were greater than 25 times the median value (criterion 2a). Another field was flagged if the pounds per unit were greater than 50 times the median (criterion 2b). The median, like the mean (average), is a measure of the location of a set of values and is defined as the value in the set that has an equal number of values above and below it. It was used rather than the mean because it is not as likely to be affected be a few extreme outliers. The median was calculated from the set of all use rates of the same pesticide product and uses as that of each record being examined. By the same uses, I mean the uses of a product on the same crop or site, same unit treated, and same record type. A record type is basically either an agricultural or non-agricultural use, but this explained more fully in Appendix I. The set of uses which have the same product, site, unit treated, and record type will be called a "use type".

Criterion 3: Pounds per unit of product is larger than the median + 10 × median deviation or the median + 50 × median deviation.

Records were flagged by criterion 3 if the pounds of a pesticide product per unit treated were greater than the median plus 10 times the median deviation (criterion 3a). Another field was flagged if the pounds per unit were greater than the median plus 50 times the median deviation (criterion 3b). As with criterion 2 the median was calculated from the set of all use rates of the same use type as that of each record being examined. The median deviation is a measure of the dispersion of a distribution, similar to the standard deviation, but based on medians rather than means. It is defined as the median of the absolute values of the differences of each record with the median.

Criterion 4: Pounds per unit of product is larger than a value generated using a neural network.

Records were flagged by criterion 4 if the pounds of a pesticide product per unit treated were greater than one to four limit values (criteria 4a, 4b, 4c, and 4d) that were calculated using a neural network procedure.

A neural network is a special kind of function that calculates a set of output values from a set of input values. This function has a large number of parameters that must be determined so that the function will give the correct outputs for every possible set of inputs. The values for these parameters are found by a procedure that involves presenting to the neural network program a set of data consisting of many sets of input and corresponding output values. The program then adjusts the parameters in the neural network function until it produces the correct output values for each input set. Once parameter values are found so that the neural network produces the correct outputs from the data it is given, it can then be used to produce appropriate output values for any input data provided to it.

The data used to train the neural network used in the PUR outlier program were generated from frequency distributions of the pounds of pesticide product per unit treated for a selected set of pesticides and sites. Groups of pesticides and sites were chosen that included a wide range of types of distributions, including many unusual distributions. Two hundred frequency distributions were plotted and then these plots were examined by 12 scientists in DPR who marked values on each plot they thought were outliers.

The results of this survey were summarized by four outlier limit values, which were used as the output values for the neural network. The input values were a set of statistical measures that described the frequency distributions. These sets of input and output values were used to find the parameter values in the neural network function. Once these parameter values were found, the neural network was ready to find the four outlier limit values for any distribution.

<u>Criterion 5: Acres treated is larger than 700.</u>

Records were flagged in the PUR by criterion 5 if the acres treated was greater than 700. A field in the PUR is defined to be an area that is no larger than a section. A section is limited to 640 acres (or slightly larger in some unusual cases).

Outlier criteria evaluation

Generally, people want to know what values are outliers so they can exclude those values from an analysis. If there are many different criteria, how can one use these criteria to decide what values to include and what to exclude? To make that decision, one needs to have a good understanding of the criteria, their advantages and disadvantages, and situations where they are likely to be most useful and least useful.

The outlier program may make mistakes either by flagging records that were not really outliers (type I error) or by not flagging records that were outliers (type II error). If you want to be conservative, in the sense of not excluding valid records from an analysis, you would want to minimize type I errors. Table 1 summarizes situations where each criterion is most appropriate and where each is sometimes not appropriate.

<u>Classifications of situations in the PUR</u>

The types of situations listed in Table 1 that are important for identifying outliers include cases where the typical rate of use is high or low (e.g. use rates are usually low for pesticides such as sex pheromones but high for pesticides such as sulfur) and where the units treated are not in acres. However, most of the situations are descriptions of types of frequency distributions. The distributions that are referred to here are the number of records (or applications) with each value of use rate for a particular use type (for some examples of these distributions see Fig. 1 in Appendix I , where this figure is explained more fully).

Normal distributions. Most people are probably aware that many properties in nature have a normal distribution, the typical bell shaped curve. Most parametric statistics are valid only if the distribution is normal (or close to normal). Most pesticide uses are not even close to being normally distributed, which is one of the main reasons why parametric statistics cannot be used to characterize outliers.

Few records. Many distributions are hard to characterize because they have too few records for a particular use type.

Broad distributions. Broad distributions have use rate values that are spread over a large range. These distributions have a large standard deviation.

Narrow distributions. Narrow distributions have most of their values close to one another. These distributions have a small standard deviation.

Many records with the same rate. There are many distributions that have a high percentage of values at, or near, the same rate. This situation is common in this database because there is often a recommended use rate for a particular pesticide product on a particular site and most people may use that rate. This kind of distribution is known as a leptokurtic distribution and has a positive kurtosis value.

Multimodal distributions. Multimodal distributions have more than one mode; that is, the distribution curve has more than one peak. This situation could occur, for example, if there are two different recommended rates and most of the uses are at or near those rates, thus creating a bimodal distribution. This kind of distribution is known as a platykurtic distribution and has a negative kurtosis value.

Comparing the criteria

Criterion 1 works for any type of distribution, regardless of the number of records or range of values (Table 1). However, if the typical use rates of a pesticide are unusually high or, criterion 1 can make either type I or type II errors. Also, because criterion 1 only applies to records with units in acres, it will miss outliers in any record measured with any other unit and so make many type II errors.

Criterion 2 is an improvement of criterion 1 in that in takes in account the typical rates of some use type and because it can be used for records treated on any unit, not just acres. However, there must be other records of the same use type so that a comparison can be made, so it can make errors if there are few records of a use type. Also it ignores the usual range of use rates, so can make errors if the distribution is very broad or very narrow.

Criterion 3 is a further improvement of criterion 2 by adding consideration of the range in values of the rates. That is, it increases the outlier limits for broad distributions and decreases it for narrow distributions, thus improving the main disadvantages of criterion

2. In other ways it is similar to criterion 2. However, criterion 3 fails for certain unusual types of distributions.

Criterion 4, using neural networks, is a completely different way of identifying outliers. This criterion worked very well in nearly all the types of situations where the other criteria failed (Table 1). The only situation where it could make an error is when there are few records of a use type.

The series of criteria from 1 through 4 become more complex, but also better in identifying outliers. This is most easily seen by comparing the list of situations where each criterion works well in the first column of Table 1. This list gets larger as one moves down from criterion 1 to criterion 4.

None of the criteria were completely satisfactory, but in general criterion 4 (using neural networks) gave the best results. Criteria 1 and 2 failed to flag many records that were obviously outliers and probably, in a few cases, flagged records that were probably not outliers. Criterion 3 worked well only for normal distributions; for most types of distributions it flagged too many valid records to be used uncritically. However, each criterion can find some outliers that the others cannot and thus it may be advantageous to use them in combination. Criteria 1 and 4, especially, appeared to be good complements.

One might wonder why not use criterion 4 and ignore the rest. For a quick analysis this is probably the easiest and best procedure. However, if one is doing an analysis for only one or a few pesticides or for a crop with only a few pesticide applications, a more careful analysis could be done by looking at other criteria, which may reveal some outliers not found by criterion 4. This is most likely to occur when there are few records of a particular use type. Using both criterion 1 and criterion 4 is better if one knows that the pesticides being analyzed are not likely to be used at such high rates that criterion 1 would erroneously flag some valid records. Also, if one has time to look carefully at many records, using the other criteria may help in making a more informed judgment about what is an outlier or not. For most situations, criterion 3 should not be used since its error rate (especially for type I errors) is high. However, even criterion 3 could be used to find some outliers not found by others if its validity is confirmed, for example, by generating the frequency distributions for each use type.

Numbers and effect of outliers in the PUR

Number of outliers found by each criterion

Total number of outliers for all of California. To examine the results of the outlier program, queries were run to summarize the number and percentages of records that were flagged by criteria 1 – 4 for each year of the PUR (Tables 2a and 2b). In addition to the values for each criterion, these tables also present the numbers and percentages of records flagged by the combination of three criteria—namely records that are flagged by either criterion 1a or by criterion 2b or by criterion 4d. Criterion 3 was not used for this criteria combination because it resulted in too many type I errors to be useful in general summary statistics. I used criteria 2b and 4d because they were the more conservative values. Criterion 1a was used rather than 1b, because 1b was too conservative (that is, so high it generated many type II errors). The combination of the 3 criteria values can be thought of as a summary of the most important criteria.

The total number of records used in Table 2 includes only records that have a positive number of units treated and does not include adjuvants. Records with 0 units treated were not included because the outlier program only checks records by the rate of application or by acreage treated and so can not provide any idea of how many of these records are outliers. Adjuvants were not included because their rates of use vary so widely and are so inconsistent that the number of outliers is not very meaningful. Also, most people are not interested in pounds of adjuvants used.

For all years, criterion 4a, the most lenient neural network criterion, flagged by far the most number of records, followed by criteria 3a and 3b. Criterion 5 (acres greater than 700) flagged the fewest which was to be expected since the Information Systems Branch already checks for this criterion. The next fewest outliers were flagged by criteria 1a and 1b. These values are consistent with the above evaluations of the criteria. One would expect few criterion 1 outliers, primarily because this criterion did not check records whose units were not in acres. Also, the limits for both criteria 1a and 1b seem too high and thus miss many outliers (type I errors). In contrast, there are many valid records that are flagged by criterion 3 (type I errors), primarily because there are many records of the same rate. The high percentages of outliers for criteria 4a and 4b (from about 2% to 13%) suggest that these criteria limits are too liberal and thus probably not as useful as the other criteria. The number of outliers for criteria 2a and 2b and 4c and 4d seem reasonable based on the sample of distributions seen in Fig. 1.

By most criteria, the percentages of records that had outliers decreased from year to year. However, the percentages increased from year to year (except for 1995) for criteria 1a and 1b and increased slightly from 1994 to 1995 for criteria 3a, 3b, 4c, and 4d.

Percentage of outliers for each county. In order to get a better understanding of where and why outliers are found, queries were run to compare percentages of records with outliers in the PUR for each county in California and for each year (Tables 3a - 3d). Based on the combination of three criteria (that is, designating a record an outlier if it violated either criteria 1a, 2b, or 4d), the counties with the highest percentages of outliers were San Mateo, Orange, Inyo, San Diego, and San Francisco (Table 3d). Among the larger pesticide using counties San Mateo, Orange, San Diego, Santa Clara, Alameda, and Los Angeles stood out in some years and by some of the criteria, having more than 1% of their records with outliers (Tables 3a - 3d). However, the high ranking of San Mateo is due mostly to an unusually high percent in 1995 according to criterion 4d. There appears to be a general tendency for primarily urban counties to have a higher percentage of outliers than agricultural counties. Otherwise there are no counties which are consistently much better or worse than others.

Percentage of outliers for each active ingredient. Queries were also run to compare percentages of outliers in the PUR for each active ingredient in California and for each year (Tables 4a - 4d). There are a dozen active ingredients with total percentages over 10% using the combination of criteria. Many of the pesticides with a high percentage of outliers are somewhat special, unusual, or used in non-agricultural sites. For example, some of the pesticides included are alcohols, sex pheromones (E-8-dodecenyl acetate, etc.), bleach (calcium hypochlorite), garlic, soap, sawdust, insect and plant hormones, and biologicals (*Bacillus thuringiensis* and *Agrobacterium radiobacter*). The sex pheromones are usually used in extremely low rates and some of the outliers are extremely large relative to these usual rates (one value was over 2 million times the median). It is possible that these extreme values were due to problems with misplaced decimals during data entry. Fumigants also appear in this category, especially using criterion 1. This is because criterion 1 has different limit values for fumigants and all other pesticides and suggests that the limit value chosen for fumigants was too low. Thus, the appearance of fumigants with high percentages of outliers using criterion 1 is misleading.

<u>Effect of removing outliers on total pounds of active ingredient reported</u>

The number of outliers found by the PUR outlier program reveals a lot about the possible sources of errors, but in general the percentage of outliers seems fairly low (in most cases much less than 1%). From this one might think that outlier errors were not a serious problem. However, even one extremely large outlier can greatly distort analyses based on summary statistics. Probably a better indication of the effect of outliers on analyses of the PUR can be found by comparing sums of active ingredients with and without outlier records included. People might be interested in summing the pounds of pesticides in many different ways, such as the sums of all active ingredients per county, sum of each active ingredient for all of California, or sum of different types of pesticides, etc. Obviously, summing the pounds used of different active ingredients could be misleading since pesticides are used at a widely differing rates, but such sums may be of interest to get a very rough estimate of pesticide use between different categories, such as between different counties.

Change in pounds of active ingredient for each county. In six counties there was more than a 5% increase in the total pounds of active ingredients used for all of 1995 if records that were identified as outliers by the criteria 1a, 2b, or 4d (Table 5d) were included. In two counties (Del Norte and Mariposa) there was more than a 10% increase in pounds of active ingredient used. These could be significant differences for some analyses.

These results, using the combined criteria, are very similar to those using only criterion 4d (Table 5c) suggesting that this criterion captures most of the large outliers. The percentage changes in active ingredient used for some counties (such as Tulare, Imperial, and San Joaquin) are similar no matter what criteria are used, while percentage changes for other counties (such as Del Norte, Contra Costa, and Madera) are quite different (Tables 5a – 5c).

There is not a strong correlation between counties with large percentage changes in pounds of active ingredients used with the counties with high percentages of outliers (Tables 3 and 5). This suggests that there are probably just a few very extreme outliers.

Change in pounds of active ingredient for each active ingredient. The effect of including outliers is dramatic when calculating the total number of pounds of active ingredient reported during 1995 (Table 6). For eight active ingredients including outliers identified by the combination of criteria increased the total pounds reported by over 1000% (Table 6d). Even including only the outliers identified by criterion 1 increased the total pounds reported for carbaryl for the state from 0.8 million pounds to nearly 1.5 million pounds (Table 6a). This change is due to a single extreme outlier value. The largest change in pounds reported, for *Agrobacterium radiobacter*, was nearly 7,000%. The presence of these outliers would seriously affect any use analysis that involved these pesticides and demonstrates the importance of identifying the outliers.

The percentage changes using criteria 2 and 4 are very similar for nearly all the 50 most affected active ingredients. The results using criterion 1 are very different because criterion 1did not identify any outliers from most of the active ingredients that had records flagged by the other criteria. I have not looked closely at all these active ingredients but the reason for most of the differences is that the active ingredients missing from the criterion 1 list are usually used at low rates. Situations where usual use rates are low often result in type II errors for criterion 1 (Table 1).

Conclusion

A computer program was developed to identify records in the PUR that had extreme rates of use. Four different types of criteria were used to identify outliers in rate of use and one criterion for acres treated. Criterion 1 flagged records if the pounds of active ingredient per acre was greater than 200 or 400 (for non-fumigants) or 1000 or 2000 (for fumigants); criterion 2 flagged records if the pounds of pesticide product per unit treated was greater than 25 or 50 times the median of all similar uses; criterion 3 flagged records if pounds of product per unit treated was greater than the median of all similar uses plus 10 or 50 times the median deviation; criterion 4 flagged records if the pounds of product per unit treated was greater than a value determined by a neural network procedure which mimicked judgments that people would make in identifying outliers; and criterion 5 flagged records if the acres treated was greater than 700.

For most situations, criterion 4 appeared to identify outliers most accurately. The main situation where it could fail to find outliers is where there are few records in a use type (that is, applications of the same pesticide product, on the same site, using the same unit treated, and same record type). For this situation, criterion 1 can be used. Although criterion 2 and, especially, criterion 3 had problems in certain situations, these criteria can still be useful in verifying the results of criterion 4 and can help one pick out outliers that criterion 4 might have missed.

Each criterion had at least two different outlier limits. Analysts can use these different limits, along with the different types of criteria, for different purposes. If a quick query is necessary, one would probably want to exclude only the most extreme, and thus the most certain errors, from the analysis. If one had more time and needed a more detailed analysis, the records flagged by less extreme criteria could be examined to determine with more confidence whether or not they were truly errors.

However, there are still more improvements that can be made in the outlier procedures. Criterion 4 could be improved by using a larger training set and by trying out different training procedures. It should also be noted that all these criteria only look at rates of use, which require a positive value for units treated. Actually, many records in the PUR have values for pounds of product used but either have no value for unit treated or a value of 0. None of these kinds of records are examined by any of these criteria.

All criteria, except criterion 5, found a significant number of outliers (from 1.3% to 13% of all the records) in each year of the PUR. Criteria 3a, 3b, 4a, and 4b identified more records than can be examined in a reasonable amount of time unless one is looking at a small subset of the data. Outliers did not appear to be especially more common in some counties than others except that urban areas tended to have more outliers than rural areas. The percentage of outliers was rarely above 5% by any of the criteria. Similarly, there was no obvious pattern to the types of chemicals with more outliers, unless it was that many of these chemicals were not typical pesticides. There were about a dozen active ingredients which had percentages of outliers greater than 10%.

However, looking at just the number or percentages of outliers does not indicate the effect outliers could have on an analysis. Some of the outlier values were quite extreme. Many use rate outliers were over 100 times the median value for their use type and one was over 2 million times the median. To determine the effect of these outliers on various kinds of analyses, one should look at the change in total number of pounds of pesticides that occurs with and without outliers present. If outliers were not removed, there would be more than a 5% over reporting of pesticide use in about 6 counties. Even more dramatic, if outliers were not removed, there would be more than a 1000% over reporting for about 8 active ingredients and more than a 10% over reporting for many more chemicals. These results illustrate the seriousness of the outlier problem in the PUR.

Table 1. Situations where each criterion usually successfully flags outliers and situations where each criterion may fail, either by flagging valid records (Type I error) or not flagging records that are outliers (Type II error). The situations are mostly different kinds of distributions in the rates of use for a use type.

	Situations where each criterion works well	**Situations where valid records may be flagged (Type I Error)**	**Situations where outlier records may not be flagged (Type II Error)**
Criterion 1 (lbs AI/acre > fixed value)	• Any type of distribution • Few records • Many records same rate	• Usual use rates high	• Usual use rates low • Units treated not in acres
Criterion 2 (lbs product/unit > 25 or 50 X median)	• Usual use rates high • Usual use rates low • Units treated not in acres • Many records same rate	• Broad distributions	• Narrow distributions • Few records
Criterion 3 (lbs product/unit > median + 10 or 50 X median deviation)	• Normal distributions • Broad distributions • Narrow distributions • Usual use rates high • Usual use rates low • Units treated not in acres	• Many records same rate • Mulitmodal distributions	• Few records
Criterion 4 (lbs product/unit > neural network limit)	• Normal distributions • Broad distributions • Narrow distributions • Usual use rates high • Usual use rates low • Units treated not in acres • Many records same rate • Mulitmodal distributions		• Few records

Table 2a. Number of outliers found by the outlier program for each of the the different criteria found in the Department of Pesticide Regulation's Pesticide Use Report (PUR) for the years 1991 through 1995. Full explanation of the criteria are given in the text. The database field name of each criteria indicates the basis of each criteria. The last row gives the total number of records in the PUR for each year in which the number of units treated was greater than 0 and in which the pesticide was not an adjuvant, that is, the number of records which were checked for outliers.

Criteria	Criteria Name	1991	1992	1993	1994	1995
124	1a or 2b or 4d	10,702	10,318	8,346	8,805	10,340
1a	ai_a_1000_200	497	574	625	759	263
1b	ai_a_2000_400	250	258	322	415	212
2a	prd_u_25m	7,700	6,102	4,677	4,197	3,599
2b	prd_u_50m	5,060	4,634	3,213	2,722	2,069
3a	prd_u_10md	58,629	58,240	66,365	66,262	71,239
3b	prd_u_50md	37,949	37,540	43,541	45,423	49,613
4a	nn1	168,376	182,339	190,131	198,641	200,172
4b	nn2	35,912	35,383	34,589	35,185	36,595
4c	nn3	16,703	16,556	15,362	15,115	16,762
4d	nn4	9,934	9,890	7,939	8,290	9,834
5	acre700	98	73	36	19	14
Number of records checked		1,296,322	1,406,688	1,498,569	1,569,480	1,670,487

Table 2b. Same data as in Table 1a, but number expressed as percentage of outliers of total number records in each year.

Criteria	Criteria Name	1991	1992	1993	1994	1995
124	1a or 2b or 4d	0.826	0.733	0.557	0.561	0.619
1a	ai_a_1000_200	0.038	0.041	0.042	0.048	0.016
1b	ai_a_2000_400	0.019	0.018	0.021	0.026	0.013
2a	prd_u_25m	0.594	0.434	0.312	0.267	0.215
2b	prd_u_50m	0.390	0.329	0.214	0.173	0.124
3a	prd_u_10md	4.523	4.140	4.429	4.222	4.265
3b	prd_u_50md	2.927	2.669	2.906	2.894	2.970
4a	nn1	12.989	12.962	12.688	12.656	11.983
4b	nn2	2.770	2.515	2.308	2.242	2.191
4c	nn3	1.288	1.177	1.025	0.963	1.003
4d	nn4	0.766	0.703	0.530	0.528	0.589
5	acre700	0.008	0.005	0.002	0.001	0.001

Table 3a. Percentage of outliers in the PUR found by the outlier program using criterion 1a for each county in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county where unit treated is greater than 0. The values are sorted by values in the total column.

County	1991	1992	1993	1994	1995	Total	Num Records
Orange	0.015	0.128	1.795	1.644	0.092	0.718	13,435
Tuolumne	0.299	0.759	0.562	0.000	0.000	0.345	290
Lassen	0.000	0.000	0.000	0.813	0.000	0.192	312
Marin	0.000	0.000	0.935	0.000	0.000	0.180	667
Calaveras	0.713	0.215	0.000	0.000	0.000	0.136	588
San Diego	0.120	0.186	0.172	0.082	0.057	0.123	56,122
El Dorado	0.050	0.130	0.126	0.249	0.000	0.119	2,188
Mariposa	0.962	0.000	0.000	0.000	0.000	0.112	178
Imperial	0.008	0.024	0.027	0.102	0.350	0.105	48,267
Nevada	0.394	0.000	0.154	0.000	0.000	0.099	604
Los Angeles	0.239	0.117	0.084	0.032	0.000	0.094	8,258
Sutter	0.083	0.083	0.046	0.112	0.066	0.078	17,013
Sacramento	0.060	0.125	0.116	0.081	0.000	0.076	11,913
Riverside	0.088	0.160	0.070	0.030	0.000	0.068	44,403
Sonoma	0.090	0.180	0.070	0.039	0.003	0.064	26,384
Placer	0.000	0.033	0.068	0.065	0.099	0.054	2,988
Yolo	0.018	0.028	0.023	0.168	0.020	0.053	17,824
Napa	0.043	0.015	0.060	0.057	0.000	0.034	14,511
Ventura	0.063	0.039	0.042	0.029	0.002	0.033	50,905
Yuba	0.086	0.039	0.020	0.028	0.000	0.033	4,795
San Joaquin	0.031	0.067	0.042	0.026	0.002	0.033	57,932
Santa Barbara	0.036	0.057	0.036	0.026	0.001	0.030	69,218
San Bernardino	0.013	0.070	0.033	0.021	0.009	0.028	9,216
Lake	0.099	0.000	0.028	0.000	0.000	0.027	5,856
San Mateo	0.034	0.012	0.069	0.016	0.000	0.026	16,792
Fresno	0.034	0.033	0.031	0.031	0.000	0.024	202,621
Colusa	0.044	0.024	0.045	0.026	0.000	0.024	13,896
Butte	0.026	0.032	0.063	0.000	0.000	0.023	22,163
Humboldt	0.000	0.000	0.154	0.000	0.000	0.022	1,799
Tulare	0.032	0.024	0.026	0.022	0.003	0.021	129,983
Madera	0.019	0.023	0.009	0.053	0.000	0.020	40,146
Mendocino	0.028	0.039	0.000	0.037	0.000	0.020	7,983
Solano	0.035	0.000	0.008	0.057	0.000	0.020	12,072
Santa Clara	0.067	0.013	0.000	0.031	0.005	0.020	16,196
Shasta	0.000	0.000	0.000	0.086	0.000	0.018	1,083
Santa Cruz	0.061	0.018	0.004	0.004	0.000	0.017	26,599
Tehama	0.000	0.051	0.017	0.000	0.014	0.017	5,784
Stanislaus	0.033	0.018	0.021	0.011	0.000	0.016	55,765
Alameda	0.000	0.000	0.000	0.070	0.000	0.016	4,972
Amador	0.000	0.000	0.079	0.000	0.000	0.014	1,390
Kings	0.019	0.018	0.031	0.009	0.000	0.014	31,725
San Luis Obispo	0.015	0.013	0.015	0.018	0.000	0.012	45,434
Modoc	0.000	0.053	0.000	0.000	0.000	0.011	1,792
Monterey	0.033	0.012	0.009	0.006	0.000	0.011	213,565
San Benito	0.009	0.006	0.006	0.011	0.006	0.007	16,230

Table 3a. Percentage of outliers in the PUR found by the outlier program using criterion 1a for each county in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county where unit treated is greater than 0. The values are sorted by values in the total column.

County	1991	1992	1993	1994	1995	Total	Num Records
Siskiyou	0.000	0.000	0.000	0.030	0.000	0.007	2,997
Kern	0.010	0.004	0.011	0.010	0.000	0.006	80,125
Contra Costa	0.000	0.000	0.025	0.000	0.000	0.005	7,579
Glenn	0.008	0.008	0.000	0.007	0.000	0.005	13,120
Merced	0.010	0.006	0.002	0.004	0.000	0.004	49,287
Alpine	0.000	0.000	0.000			0.000	3
Del Norte	0.000	0.000	0.000	0.000	0.000	0.000	2,964
Inyo	0.000	0.000	0.000	0.000	0.000	0.000	72
Mono	0.000	0.000	0.000	0.000	0.000	0.000	66
Plumas	0.000	0.000	0.000	0.000		0.000	45
San Francisco	0.000	0.000	0.000	0.000	0.000	0.000	48
Sierra	0.000	0.000	0.000	0.000		0.000	10
Trinity	0.000	0.000	0.000	0.000	0.000	0.000	135

Table 3b. Percentage of outliers in the PUR found by the outlier program using criterion 2b for each county in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county where unit treated is greater than 0. The values are sorted by values in the total column.

County	1991	1992	1993	1994	1995	Total	Num Records
Inyo	4.839	4.762	2.586	0.000	0.000	2.793	72
San Francisco	2.857	5.172	2.381	0.000	0.000	2.101	48
Orange	0.615	0.612	4.179	3.092	1.272	1.897	13,435
San Diego	3.045	4.231	1.075	0.430	0.380	1.809	56,122
Tuolumne	1.198	4.557	0.000	0.000	0.000	1.516	290
San Benito	4.518	0.077	0.256	0.159	0.155	0.789	16,230
Alameda	0.341	0.708	0.466	0.907	1.160	0.688	4,972
Mariposa	0.962	1.274	0.000	0.546	0.897	0.675	178
Los Angeles	0.767	0.422	0.312	0.800	0.578	0.579	8,258
San Mateo	0.742	0.533	0.482	0.196	0.883	0.556	16,792
Marin	0.562	0.839	0.779	0.238	0.331	0.539	667
Sonoma	0.279	2.060	0.563	0.208	0.174	0.537	26,384
Santa Cruz	0.773	0.219	1.532	0.102	0.074	0.533	26,599
Plumas	0.000	0.000	0.000	1.613		0.446	45
Humboldt	0.100	0.321	0.770	0.118	0.843	0.422	1,799
Calaveras	1.425	0.429	0.000	0.000	0.584	0.408	588
San Bernardino	0.378	0.361	0.445	0.206	0.551	0.393	9,216
Santa Clara	0.382	0.324	0.602	0.343	0.316	0.388	16,196
El Dorado	0.656	0.261	0.252	0.290	0.107	0.311	2,188
Riverside	0.244	0.411	0.381	0.303	0.167	0.299	44,403
San Joaquin	0.352	0.217	0.206	0.201	0.118	0.213	57,932
Ventura	0.594	0.137	0.117	0.111	0.132	0.207	50,905
Sacramento	0.111	0.150	0.248	0.324	0.157	0.196	11,913
Santa Barbara	0.198	0.280	0.192	0.139	0.142	0.187	69,218
Del Norte	0.369	0.338	0.084	0.187	0.031	0.182	2,964
Placer	0.143	0.131	0.270	0.065	0.297	0.181	2,988
Contra Costa	0.408	0.251	0.063	0.097	0.089	0.172	7,579
Nevada	0.591	0.327	0.000	0.000	0.000	0.166	604
Butte	0.148	0.134	0.146	0.204	0.113	0.150	22,163
Trinity	0.000	0.000	0.529	0.000	0.000	0.149	135
Mendocino	0.127	0.261	0.156	0.086	0.058	0.135	7,983
Lassen	0.341	0.000	0.000	0.000	0.377	0.128	312
Sutter	0.131	0.094	0.115	0.095	0.197	0.127	17,013
Monterey	0.475	0.060	0.053	0.065	0.044	0.126	213,565
Fresno	0.148	0.119	0.128	0.138	0.086	0.121	202,621
Lake	0.231	0.048	0.085	0.066	0.131	0.116	5,856
Amador	0.094	0.084	0.238	0.096	0.075	0.115	1,390
Solano	0.218	0.154	0.058	0.057	0.024	0.101	12,072
Yolo	0.142	0.158	0.087	0.043	0.066	0.098	17,824
San Luis Obispo	0.040	0.056	0.077	0.182	0.103	0.095	45,434
Yuba	0.173	0.058	0.039	0.112	0.091	0.092	4,795
Colusa	0.144	0.110	0.090	0.066	0.052	0.085	13,896
Madera	0.094	0.081	0.071	0.095	0.077	0.083	40,146
Stanislaus	0.089	0.096	0.077	0.097	0.040	0.080	55,765
Kings	0.066	0.084	0.131	0.061	0.058	0.078	31,725
Tulare	0.095	0.075	0.089	0.079	0.051	0.077	129,983

Table 3b. Percentage of outliers in the PUR found by the outlier program using criterion 2b for each county in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county where unit treated is greater than 0. The values are sorted by values in the total column.

County	1991	1992	1993	1994	1995	Total	Num Records
Merced	0.099	0.070	0.065	0.098	0.036	0.071	49,287
Modoc	0.000	0.105	0.099	0.000	0.093	0.067	1,792
Kern	0.071	0.063	0.073	0.081	0.048	0.066	80,125
Tehama	0.057	0.051	0.084	0.063	0.057	0.062	5,784
Siskiyou	0.162	0.036	0.000	0.089	0.030	0.060	2,997
Napa	0.104	0.030	0.054	0.092	0.018	0.057	14,511
Glenn	0.105	0.057	0.047	0.021	0.044	0.053	13,120
Imperial	0.016	0.067	0.084	0.046	0.051	0.049	48,267
Shasta	0.094	0.000	0.000	0.086	0.000	0.037	1,083
Alpine	0.000	0.000	0.000			0.000	3
Mono	0.000	0.000	0.000	0.000	0.000	0.000	66
Sierra	0.000	0.000	0.000	0.000		0.000	10

Table 3c. Percentage of outliers in the PUR found by the outlier program using criterion 4d for each county in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county where unit treated is greater than 0. The values are sorted by values in the total column.

County	1991	1992	1993	1994	1995	Total	Num Records
San Mateo	2.782	3.155	1.233	0.796	13.437	4.271	16,792
Inyo	4.839	4.762	2.586	0.000	0.000	2.793	72
Orange	1.169	1.306	4.608	3.670	2.544	2.683	13,435
San Diego	3.978	5.114	1.893	0.798	1.157	2.560	56,122
San Francisco	5.714	5.172	2.381	0.000	0.000	2.521	48
Santa Clara	1.003	2.313	1.237	3.539	1.145	1.940	16,196
Tuolumne	1.198	4.557	3.371	0.000	0.000	1.930	290
Mariposa	0.962	3.185	0.450	4.372	0.897	1.912	178
Alameda	1.248	1.682	1.179	2.617	2.544	1.846	4,972
Calaveras	2.138	0.858	2.830	2.052	0.876	1.769	588
Los Angeles	1.963	1.184	1.127	1.621	2.723	1.686	8,258
Shasta	0.843	0.646	1.604	1.468	2.518	1.441	1,083
Humboldt	1.403	0.962	1.232	1.948	1.311	1.367	1,799
Trinity	3.247	0.000	0.529	0.800	0.000	1.040	135
San Bernardino	1.096	0.477	0.728	1.666	1.029	1.011	9,216
Sonoma	0.794	2.580	0.916	0.417	0.753	0.959	26,384
Santa Cruz	1.062	0.523	1.848	0.326	1.004	0.940	26,599
Ventura	1.242	1.016	0.684	0.861	0.755	0.901	50,905
Plumas	0.000	0.000	1.887	1.613		0.893	45
Riverside	0.759	1.072	0.973	0.859	0.740	0.878	44,403
Contra Costa	1.346	0.516	0.325	0.752	0.953	0.763	7,579
Marin	0.749	0.839	1.558	0.357	0.331	0.749	667
Nevada	0.787	0.491	0.770	1.178	0.175	0.696	604
San Benito	2.681	0.172	0.281	0.478	0.359	0.668	16,230
Mono	0.000	0.000	1.042	1.176	0.000	0.604	66
Del Norte	0.533	1.098	0.422	0.404	0.617	0.587	2,964
Yolo	1.649	0.492	0.227	0.429	0.240	0.583	17,824
Amador	0.566	0.168	0.397	1.342	0.224	0.518	1,390
Placer	1.069	0.394	0.371	0.421	0.362	0.515	2,988
Fresno	0.677	0.548	0.489	0.405	0.450	0.502	202,621
Sacramento	0.412	0.342	0.355	0.775	0.621	0.499	11,913
Santa Barbara	0.436	0.365	0.597	0.389	0.602	0.483	69,218
Madera	0.687	0.384	0.369	0.566	0.432	0.482	40,146
El Dorado	0.807	0.478	0.378	0.332	0.429	0.475	2,188
Mendocino	0.452	0.628	0.469	0.440	0.357	0.466	7,983
Yuba	0.863	0.405	0.294	0.279	0.329	0.434	4,795
San Luis Obispo	0.299	0.329	0.497	0.547	0.391	0.418	45,434
San Joaquin	0.444	0.451	0.480	0.369	0.294	0.404	57,932
Stanislaus	0.258	0.382	0.358	0.638	0.312	0.398	55,765
Sutter	0.316	0.533	0.375	0.236	0.493	0.396	17,013
Lake	0.248	0.511	0.566	0.132	0.291	0.379	5,856
Butte	0.424	0.360	0.344	0.446	0.280	0.377	22,163
Kern	0.393	0.322	0.373	0.331	0.332	0.347	80,125
Colusa	0.389	0.408	0.234	0.351	0.284	0.325	13,896
Tulare	0.487	0.332	0.276	0.313	0.234	0.320	129,983
Solano	0.427	0.236	0.248	0.506	0.180	0.318	12,072

Table 3c. Percentage of outliers in the PUR found by the outlier program using criterion 4d for each county in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county where unit treated is greater than 0. The values are sorted by values in the total column.

County	1991	1992	1993	1994	1995	Total	Num Records
Tehama	0.209	0.376	0.420	0.126	0.382	0.315	5,784
Napa	0.347	0.207	0.355	0.389	0.257	0.310	14,511
Monterey	0.642	0.208	0.215	0.255	0.185	0.288	213,565
Glenn	0.258	0.378	0.241	0.077	0.356	0.258	13,120
Lassen	0.341	0.276	0.000	0.271	0.377	0.256	312
Merced	0.298	0.199	0.266	0.326	0.194	0.254	49,287
Kings	0.194	0.276	0.324	0.190	0.152	0.220	31,725
Modoc	0.323	0.210	0.099	0.445	0.093	0.212	1,792
Imperial	0.186	0.195	0.216	0.295	0.169	0.210	48,267
Siskiyou	0.283	0.572	0.034	0.178	0.030	0.207	2,997
Alpine	0.000	0.000	0.000			0.000	3
Sierra	0.000	0.000	0.000	0.000		0.000	10

Table 3d. Percentage of outliers in the PUR found by the outlier program using criteria 1a, 2b, or 4d for each county in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county where unit treated is greater than 0. The values are sorted by values in the total column.

County	1991	1992	1993	1994	1995	Total	Num Records
San Mateo	2.858	3.155	1.279	0.802	13.524	4.313	16,792
Orange	1.246	1.427	4.782	3.962	2.616	2.839	13,435
Inyo	4.839	4.762	2.586	0.000	0.000	2.793	72
San Diego	4.046	5.188	2.017	0.885	1.254	2.651	56,122
San Francisco	5.714	5.172	2.381	0.000	0.000	2.521	48
Tuolumne	2.096	4.810	3.371	0.000	0.000	2.205	290
Santa Clara	1.012	2.313	1.275	3.549	1.165	1.956	16,196
Mariposa	0.962	3.185	0.450	4.372	0.897	1.912	178
Alameda	1.248	1.682	1.179	2.652	2.566	1.858	4,972
Calaveras	2.375	0.858	2.830	2.052	0.876	1.803	588
Los Angeles	2.025	1.242	1.199	1.674	2.737	1.739	8,258
Shasta	0.843	0.646	1.604	1.468	2.518	1.441	1,083
Humboldt	1.403	0.962	1.309	1.948	1.311	1.378	1,799
San Bernardino	1.122	0.558	0.923	1.687	1.387	1.159	9,216
Sonoma	0.842	2.826	0.994	0.589	0.846	1.083	26,384
Trinity	3.247	0.000	0.529	0.800	0.000	1.040	135
Santa Cruz	1.329	0.534	1.890	0.359	1.004	1.010	26,599
San Benito	4.714	0.172	0.281	0.483	0.364	0.964	16,230
Ventura	1.438	1.052	0.702	0.882	0.759	0.951	50,905
Riverside	0.834	1.176	1.073	0.889	0.766	0.944	44,403
Marin	0.749	0.839	2.492	0.357	0.331	0.929	667
Plumas	0.000	0.000	1.887	1.613		0.893	45
Contra Costa	1.346	0.530	0.338	0.752	0.953	0.768	7,579
Nevada	0.787	0.491	0.924	1.178	0.175	0.729	604
El Dorado	1.060	0.478	0.631	0.580	0.429	0.631	2,188
Yolo	1.698	0.503	0.239	0.511	0.255	0.617	17,824
Mono	0.000	0.000	1.042	1.176	0.000	0.604	66
Del Norte	0.574	1.098	0.422	0.436	0.617	0.601	2,964
Sacramento	0.446	0.459	0.562	0.793	0.621	0.574	11,913
Placer	1.069	0.394	0.405	0.421	0.428	0.536	2,988
Santa Barbara	0.513	0.500	0.612	0.402	0.618	0.531	69,218
Fresno	0.688	0.567	0.506	0.434	0.468	0.522	202,621
Amador	0.566	0.168	0.397	1.342	0.224	0.518	1,390
Madera	0.687	0.394	0.372	0.596	0.432	0.490	40,146
San Joaquin	0.636	0.525	0.512	0.425	0.353	0.482	57,932
Mendocino	0.466	0.641	0.469	0.453	0.357	0.473	7,983
Yuba	0.885	0.424	0.313	0.391	0.348	0.467	4,795
Sutter	0.337	0.577	0.404	0.354	0.586	0.458	17,013
Lassen	0.341	0.276	0.000	1.084	0.377	0.448	312
San Luis Obispo	0.307	0.337	0.510	0.566	0.396	0.429	45,434
Stanislaus	0.311	0.388	0.359	0.648	0.328	0.414	55,765
Butte	0.434	0.365	0.396	0.483	0.295	0.400	22,163
Lake	0.264	0.511	0.566	0.132	0.291	0.382	5,856
Kern	0.396	0.326	0.400	0.359	0.332	0.360	80,125
Tulare	0.503	0.346	0.298	0.326	0.244	0.335	129,983
Colusa	0.411	0.408	0.256	0.357	0.284	0.334	13,896

Table 3d. Percentage of outliers in the PUR found by the outlier program using criteria 1a, 2b, or 4d for each county in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county where unit treated is greater than 0. The values are sorted by values in the total column.

County	1991	1992	1993	1994	1995	Total	Num Records
Tehama	0.209	0.411	0.453	0.126	0.396	0.332	5,784
Solano	0.435	0.252	0.248	0.514	0.180	0.325	12,072
Napa	0.347	0.207	0.361	0.389	0.257	0.311	14,511
Imperial	0.191	0.195	0.223	0.395	0.517	0.306	48,267
Monterey	0.648	0.213	0.216	0.260	0.187	0.292	213,565
Merced	0.328	0.204	0.276	0.338	0.197	0.265	49,287
Glenn	0.258	0.386	0.241	0.083	0.356	0.261	13,120
Kings	0.194	0.276	0.337	0.193	0.152	0.223	31,725
Modoc	0.323	0.210	0.099	0.445	0.093	0.212	1,792
Siskiyou	0.283	0.572	0.034	0.178	0.030	0.207	2,997
Alpine	0.000	0.000	0.000			0.000	3
Sierra	0.000	0.000	0.000	0.000		0.000	10

Table 4a. Percentage of outliers in the PUR found by the outlier program using criterion 1a for different active ingredients in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county. The values are sorted by values in the total column, only the top 50 Ais are shown and only Ais with more than 10 records.

AI	1991	1992	1993	1994	1995	Total	Num Records
DISODIUM OCTABORATE TETRAHYDRATE			0.00	11.11	0.00	5.45	11
E-8-DODECENYL ACETATE	16.28	7.53	0.00	6.87	0.00	4.50	191
Z-8-DODECENOL	16.28	7.53	0.00	6.87	0.00	4.50	191
Z-8-DODECENYL ACETATE	16.28	7.53	0.00	6.87	0.00	4.50	191
LAURYL ALCOHOL		4.00	2.08	1.00	0.00	1.51	66
MYRISTYL ALCOHOL		4.00	2.08	1.00	0.00	1.51	66
E,E-8,10-DODECADIEN-1-OL		4.00	2.08	0.93	0.00	1.32	76
DAZOMET	3.23	6.25	0.00	0.00	0.00	0.92	44
ACROLEIN	0.00	0.00	2.63	0.00	1.28	0.88	45
ARSENIC PENTOXIDE	0.00	7.14	0.00	0.00	0.00	0.83	24
CHROMIC ACID	0.00	7.14	0.00	0.00	0.00	0.83	24
SODIUM TETRATHIOCARBONATE	0.00	25.00	0.31	1.54	0.00	0.64	126
PETROLEUM DISTILLATES, REFINED	0.00	0.58	1.33	0.62	0.19	0.60	598
PHOSPHAMIDON	1.07	0.00	0.00	0.00	0.00	0.59	102
PHOSPHAMIDON, OTHER RELATED	1.07	0.00	0.00	0.00	0.00	0.59	102
METHYL BROMIDE	0.93	0.80	0.79	0.05	0.18	0.56	8,225
MINERAL OIL	0.71	1.78	0.89	0.25	0.02	0.55	2,860
CHLOROPICRIN	0.90	0.57	0.50	0.03	0.14	0.41	3,222
DINOCAP	0.00	0.00	0.00	2.86	0.00	0.35	57
SODIUM HYPOCHLORITE	1.70	0.76	0.00	0.00	0.00	0.29	407
PETROLEUM OIL, UNCLASSIFIED	0.26	0.30	0.42	0.35	0.08	0.28	23,214
POTASH SOAP	0.21	0.31	0.30	0.27	0.05	0.22	7,693
COPPER SULFATE (PENTAHYDRATE)	0.17	0.30	0.28	0.38	0.00	0.22	3,650
DAMINOZIDE	0.22	0.81	0.05	0.03	0.00	0.22	3,972
2,4-D, BUTOXYETHANOL ESTER	0.00	0.00	1.03	0.00	0.00	0.20	590
PETROLEUM HYDROCARBONS	0.32	0.00	0.29	0.00	0.00	0.20	715
DIQUAT DIBROMIDE	0.16	0.00	0.28	0.41	0.00	0.17	2,599
SULFUR DIOXIDE	0.00	0.00	0.00	0.60	0.00	0.14	279
METHOPRENE	0.00	0.00	0.85	0.00	0.00	0.13	156
THIOPHANATE-METHYL	0.06	0.08	0.13	0.32	0.00	0.12	9,902
PCNB	0.17	0.17	0.14	0.09	0.00	0.11	2,321
ALKYL(68%C12, 32%C14)DIMETHYL ETHYLBENZYL AMMC	0.48	0.00	0.26	0.00	0.00	0.11	370
ALKYL(60%C14,30%C16,5%C12,5%C18)DIMETHYL BENZYI	0.48	0.00	0.26	0.00	0.00	0.10	384
CYCLOATE	0.16	0.19	0.00	0.16	0.00	0.10	611
MANCOZEB	0.00	0.04	0.13	0.32	0.00	0.10	9,897
RESMETHRIN, OTHER RELATED	0.48	0.00	0.00	0.00	0.00	0.10	208
OXYCARBOXIN	0.19	0.00	0.00	0.00	0.00	0.09	217
TERRAZOLE	0.16	0.00	0.00	0.00	0.00	0.09	660

LIME-SULFUR	0.33	0.10	0.08	0.00	0.00	0.09	1,108
PHOSPHOROUS	0.00	0.00	0.56	0.00	0.00	0.09	230
SAWDUST	0.00	0.00	0.56	0.00	0.00	0.09	230
SODIUM NITRATE	0.00	0.00	0.55	0.00	0.00	0.09	233
ACEPHATE	0.02	0.03	0.13	0.25	0.00	0.09	26,902
METHIOCARB	0.00	0.00	0.29	0.18	0.00	0.09	1,172
CARBON	0.00	0.00	0.55	0.00	0.00	0.08	236
BENDIOCARB	0.00	0.25	0.00	0.00	0.00	0.08	739
6-METHYL-1,3-DITHIOLO(4,5-B)QUINOXALIN-2-ONE	0.00	0.00	0.44	0.00	0.00	0.07	562
NICOSULFURON				0.21	0.00	0.07	292
BACILLUS THURINGIENSIS	0.00	0.26	0.00	0.00	0.00	0.07	607
1-BROMO-3-CHLORO-5,5-DIMETHYLHYDANTOIN	0.26	0.00	0.00	0.00	0.00	0.06	315

Table 4b. Percentage of outliers in the PUR found by the outlier program using criterion 2b for different active ingredients in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county. The values are sorted by values in the total column, only the top 50 Ais are shown and only Ais with more than 10 records.

AI	1991	1992	1993	1994	1995	Total	Num Records
LAURYL ALCOHOL		14.00	26.04	12.00	4.71	14.50	66
MYRISTYL ALCOHOL		14.00	26.04	12.00	4.71	14.50	66
E-8-DODECENYL ACETATE	20.93	17.81	3.73	13.74	14.33	13.51	191
Z-8-DODECENOL	20.93	17.81	3.73	13.74	14.33	13.51	191
Z-8-DODECENYL ACETATE	20.93	17.81	3.73	13.74	14.33	13.51	191
E,E-8,10-DODECADIEN-1-OL		14.00	26.04	11.21	4.00	12.96	76
1-BROMO-3-CHLORO-5,5-DIMETHYLHYDANTOIN	11.66	17.09	0.78	0.37	3.23	8.58	315
NONANOIC ACID					6.45	6.45	37
NONANOIC ACID, OTHER RELATED					6.45	6.45	37
ORTHO-BENZYL-PARA-CHLOROPHENOL, POTASSIUM S	0.00	0.00	13.77	0.00	0.00	4.99	92
ORTHO-PHENYLPHENOL, POTASSIUM SALT	0.00	0.00	13.77	0.00	0.00	4.99	92
PARA-TERT-AMYLPHENOL, POTASSIUM SALT	0.00	0.00	13.77	0.00	0.00	4.99	92
AGROBACTERIUM RADIOBACTER	3.92	9.30	9.09	0.00	2.33	4.76	42
ALKYL(60%C14,30%C16,5%C12,5%C18)DIMETHYL BENZ	2.86	6.93	5.41	1.85	2.36	3.59	384
SULFUR DIOXIDE	5.71	2.07	5.93	2.39	2.50	3.52	279
SULFOTEP	6.37	6.07	1.92	1.69	1.12	3.49	493
ALKYL(68%C12, 32%C14)DIMETHYL ETHYLBENZYL AMM	2.88	7.19	5.03	1.38	2.54	3.46	370
NONYLPHENOXYPOLYOXYETHYLENE ETHANOL-IODINI	0.00	8.82	0.00	0.00	0.00	2.94	20
NAA	3.85	0.00	6.73	3.42	0.98	2.93	89
PROPYLENE OXIDE	0.00	0.00	2.27	0.00	8.57	2.52	32
CHLORINE	2.76	1.55	0.48	2.56	3.95	2.35	230
DAZOMET	0.00	0.00	0.00	0.00	4.35	2.29	44
RESMETHRIN, OTHER RELATED	5.80	4.49	0.00	0.00	0.00	2.22	208
CHLORPROPHAM	0.00	3.17	2.04	0.00	6.90	2.19	55
SODIUM HYPOCHLORITE	0.00	0.76	2.74	3.51	1.89	2.11	407
CHLORMEQUAT CHLORIDE	5.28	2.96	0.73	0.85	0.45	2.09	1,590
TERRAZOLE	2.83	1.53	0.32	0.00	0.82	2.03	660
ORTHO-BENZYL-PARA-CHLOROPHENOL, SODIUM SALT	0.00	6.90	1.47	0.00	0.00	1.96	51
PARA-TERT-AMYLPHENOL, SODIUM SALT	0.00	6.78	1.47	0.00	0.00	1.95	51
TETRACHLORVINPHOS	1.21	1.63	3.45	1.16	2.14	1.89	212
RESMETHRIN	4.48	3.91	0.45	0.20	0.24	1.88	457
IMAZALIL	2.65	1.15	2.09	1.44	2.43	1.87	214
ALUMINUM PHOSPHIDE	1.67	1.32	2.05	2.27	1.69	1.80	3,442
CALCIUM HYPOCHLORITE	8.38	0.04	0.04	0.03	0.05	1.79	6,192
SOAP	0.00	0.00	0.00	3.39	0.00	1.79	22
BACILLUS THURINGIENSIS (BERLINER), SUBSP. ISRAEL	2.32	3.91	1.72	1.50	0.44	1.75	684
KINOPRENE	2.08	3.88	0.81	1.26	0.57	1.66	2,013

SULFAQUINOXALINE	6.25	0.00	0.00	0.00	0.00	1.64	12
DIENOCHLOR	2.86	2.83	2.37	0.46	0.39	1.62	5,150
IBA	0.99	1.41	2.38	1.88	0.97	1.59	465
TAU FLUVALINATE	2.98	2.65	1.13	1.02	0.52	1.55	7,564
WARFARIN	5.56	0.00	0.00	0.00	0.00	1.54	13
DDVP, OTHER RELATED	1.12	1.43	3.24	0.66	1.69	1.53	261
DDVP	1.12	1.43	3.24	0.66	1.69	1.53	261
6-METHYL-1,3-DITHIOLO(4,5-B)QUINOXALIN-2-ONE	1.01	0.71	2.86	1.34	2.07	1.49	562
NICOTINE	1.11	2.56	2.29	0.24	0.51	1.44	665
DIOCTYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	0.57	0.00	3.81	1.40	185
OCTYL DECYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	0.57	0.00	3.81	1.40	185
DIDECYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	0.57	0.00	3.81	1.40	186
ALKYL(50%C14,40%C12,10%C16)DIMETHYLBENZYL AM	0.59	1.60	0.55	0.00	3.67	1.36	191

Table 4c. Percentage of outliers in the PUR found by the outlier program using criterion 4d for different active ingredients in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county. The values are sorted by values in the total column, only the top 50 Ais are shown and only Ais with more than 10 records.

AI	1991	1992	1993	1994	1995	Total	Num Records
E-8-DODECENYL ACETATE	20.93	10.96	3.73	11.83	14.33	11.94	191
Z-8-DODECENOL	20.93	10.96	3.73	11.83	14.33	11.94	191
Z-8-DODECENYL ACETATE	20.93	10.96	3.73	11.83	14.33	11.94	191
OCTYL BICYCLOHEPTENEDICARBOXIMIDE	1.35	17.59	20.31	16.38	0.00	11.87	88
ORTHO-PHENYLPHENOL	0.00	0.00	0.00	37.50	0.00	11.65	41
METHOPRENE	1.97	5.23	14.41	17.45	15.05	10.93	156
IMAZALIL	2.65	1.15	2.09	23.44	21.46	10.57	214
FOSAMINE, AMMONIUM SALT	0.00	0.00	0.00	88.89	0.00	10.39	15
CALCIUM HYPOCHLORITE	11.03	5.40	0.46	0.09	29.84	9.76	6,192
ORTHO-PHENYLPHENOL, SODIUM SALT	0.00	8.20	3.38	18.43	14.61	9.72	193
GARLIC	0.00	0.00	0.00	17.35	1.23	9.37	124
E,E-8,10-DODECADIEN-1-OL		12.00	7.29	10.28	8.80	9.26	76
PETROLEUM DERIVATIVE RESIN	17.92	2.63	6.45	0.00	0.00	9.15	57
BACILLUS THURINGIENSIS (BERLINER), SUBSP. KURSTAKI				0.00	15.38	9.02	164
SOAP	0.00	0.00	0.00	3.39	50.00	8.93	22
LAURYL ALCOHOL		12.00	7.29	11.00	4.71	8.46	66
MYRISTYL ALCOHOL		12.00	7.29	11.00	4.71	8.46	66
PETROLEUM DISTILLATES	8.23	3.94	17.58	0.00	12.50	8.15	118
NOREA	14.57	13.84	0.00	0.00	0.00	7.93	129
NOREA, OTHER RELATED	14.57	13.84	0.00	0.00	0.00	7.93	129
2,4-D, ISOOCTYL ESTER	0.00	0.00	0.00	0.00	20.59	7.69	36
NAA	3.85	1.05	6.73	16.24	0.98	6.53	89
NONANOIC ACID					6.45	6.45	37
NONANOIC ACID, OTHER RELATED					6.45	6.45	37
8-DODECENE-1-OL, OTHER RELATED	0.00	0.00	92.31		3.03	5.96	44
TETRACHLORVINPHOS	1.21	1.63	22.17	1.16	1.60	5.38	212
ALKYL(60%C14,30%C16,5%C12,5%C18)DIMETHYL BENZYL	5.71	8.91	5.41	5.04	2.83	5.31	384
SODIUM HYPOCHLORITE	0.00	1.01	2.95	10.14	7.37	5.16	407
PETROLEUM HYDROCARBONS	15.97	2.90	1.17	1.99	0.00	5.15	715
ALKYL(68%C12, 32%C14)DIMETHYL ETHYLBENZYL AMMON	5.77	9.25	4.50	4.65	3.05	5.13	370
ORTHO-BENZYL-PARA-CHLOROPHENOL, POTASSIUM SAL	0.00	0.00	13.77	0.00	0.00	4.99	92
ORTHO-PHENYLPHENOL, POTASSIUM SALT	0.00	0.00	13.77	0.00	0.00	4.99	92
PARA-TERT-AMYLPHENOL, POTASSIUM SALT	0.00	0.00	13.77	0.00	0.00	4.99	92
1-BROMO-3-CHLORO-5,5-DIMETHYLHYDANTOIN	1.55	4.43	10.12	2.96	8.60	4.90	315
DIOCTYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	2.84	0.00	17.14	4.85	185
OCTYL DECYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	2.84	0.00	17.14	4.85	185
DIDECYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	2.84	0.00	17.14	4.85	186

AGROBACTERIUM RADIOBACTER	3.92	9.30	9.09	0.00	2.33	4.76	42
ALKYL(50%C14,40%C12,10%C16)DIMETHYLBENZYL AMMOI	0.59	1.60	2.76	0.00	16.51	4.70	191
SULFOTEP	8.30	6.07	2.40	4.23	1.12	4.63	493
CHLORINE	2.76	2.33	6.67	5.13	5.59	4.60	230
IBA	12.38	1.69	2.38	6.41	3.54	4.56	465
RESMETHRIN	6.73	4.32	5.67	0.41	3.78	4.11	457
4-AMINOPYRIDINE	10.53	0.00	0.00	0.00	0.00	3.88	21
RESMETHRIN, OTHER RELATED	10.63	4.90	2.40	0.47	0.00	3.85	208
CUBE EXTRACTS	5.26	0.00	0.00	0.00	0.00	3.77	21
SULFUR DIOXIDE	7.14	2.37	4.44	2.09	2.50	3.45	279
GIBBERELLINS	4.60	3.34	4.20	1.17	3.25	3.28	11,201
CHLOROPHACINONE	1.39	2.18	1.20	3.87	9.34	3.24	921
THIABENDAZOLE	2.92	8.58	2.79	2.13	0.85	3.22	1,851

Table 4d. Percentage of outliers in the PUR found by the outlier program using criteria 1a, 2b, or 4d for different active ingredients in California for the years 1991 through 1995. This analysis only included data on non-adjuvant pesticides. The values in the total column are the percentages over all five years. The values in the last column are the yearly mean number of records in the PUR for each county. The values are sorted by values in the total column, only the top 50 Ais are shown and only Ais with more than 10 records.

AI	1991	1992	1993	1994	1995	Total	Num Records
E,E-8,10-DODECADIEN-1-OL		14.00	26.04	11.21	8.80	14.55	76
LAURYL ALCOHOL		14.00	26.04	12.00	4.71	14.50	66
MYRISTYL ALCOHOL		14.00	26.04	12.00	4.71	14.50	66
E-8-DODECENYL ACETATE	20.93	17.81	3.73	13.74	14.33	13.51	191
Z-8-DODECENOL	20.93	17.81	3.73	13.74	14.33	13.51	191
Z-8-DODECENYL ACETATE	20.93	17.81	3.73	13.74	14.33	13.51	191
1-BROMO-3-CHLORO-5,5-DIMETHYLHYDANTOIN	11.92	21.31	10.12	2.96	11.29	12.84	315
OCTYL BICYCLOHEPTENEDICARBOXIMIDE	1.35	17.59	20.31	16.38	0.00	11.87	88
ORTHO-PHENYLPHENOL	0.00	0.00	0.00	37.50	0.00	11.65	41
METHOPRENE	1.97	5.23	15.25	17.45	15.05	11.05	156
IMAZALIL	2.65	1.15	2.09	23.92	21.46	10.66	214
FOSAMINE, AMMONIUM SALT	0.00	0.00	0.00	88.89	0.00	10.39	15
CALCIUM HYPOCHLORITE	11.89	5.40	0.46	0.09	29.84	9.94	6,192
ORTHO-PHENYLPHENOL, SODIUM SALT	0.00	8.61	3.38	18.43	14.61	9.82	193
GARLIC	0.00	0.00	0.00	17.35	1.23	9.37	124
PETROLEUM DERIVATIVE RESIN	17.92	2.63	6.45	0.00	0.00	9.15	57
BACILLUS THURINGIENSIS (BERLINER), SUBSP. KURST				0.00	15.38	9.02	164
SOAP	0.00	0.00	0.00	3.39	50.00	8.93	22
PETROLEUM DISTILLATES	8.23	3.94	17.58	0.00	12.50	8.15	118
NOREA	14.57	13.84	0.00	0.00	0.00	7.93	129
NOREA, OTHER RELATED	14.57	13.84	0.00	0.00	0.00	7.93	129
2,4-D, ISOOCTYL ESTER	0.00	0.00	0.00	0.00	20.59	7.69	36
NAA	3.85	1.05	6.73	16.24	0.98	6.53	89
NONANOIC ACID					6.45	6.45	37
NONANOIC ACID, OTHER RELATED					6.45	6.45	37
8-DODECENE-1-OL, OTHER RELATED	0.00	0.00	92.31		3.03	5.96	44
ALKYL(60%C14,30%C16,5%C12,5%C18)DIMETHYL BEN	6.19	8.91	6.19	5.04	3.30	5.63	384
SODIUM HYPOCHLORITE	1.70	1.77	2.95	10.33	7.37	5.51	407
TETRACHLORVINPHOS	1.21	1.63	22.17	1.16	2.14	5.48	212
DISODIUM OCTABORATE TETRAHYDRATE			0.00	11.11	0.00	5.45	11
ALKYL(68%C12, 32%C14)DIMETHYL ETHYLBENZYL AM	6.25	9.25	5.03	4.65	3.56	5.40	370
PETROLEUM HYDROCARBONS	16.29	2.90	1.46	1.99	0.00	5.34	715
ORTHO-BENZYL-PARA-CHLOROPHENOL, POTASSIUM	0.00	0.00	13.77	0.00	0.00	4.99	92
ORTHO-PHENYLPHENOL, POTASSIUM SALT	0.00	0.00	13.77	0.00	0.00	4.99	92
PARA-TERT-AMYLPHENOL, POTASSIUM SALT	0.00	0.00	13.77	0.00	0.00	4.99	92
DIOCTYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	2.84	0.00	17.14	4.85	185
OCTYL DECYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	2.84	0.00	17.14	4.85	185

DIDECYL DIMETHYL AMMONIUM CHLORIDE	0.60	1.60	2.84	0.00	17.14	4.85	186
SULFOTEP	8.49	6.07	2.72	4.40	1.12	4.79	493
AGROBACTERIUM RADIOBACTER	3.92	9.30	9.09	0.00	2.33	4.76	42
ALKYL(50%C14,40%C12,10%C16)DIMETHYLBENZYL AM	0.59	1.60	2.76	0.00	16.51	4.70	191
CHLORINE	2.76	2.33	6.67	5.13	5.59	4.60	230
IBA	12.38	1.69	2.38	6.41	3.54	4.56	465
RESMETHRIN	6.95	4.32	5.67	0.41	3.78	4.16	457
RESMETHRIN, OTHER RELATED	11.11	4.90	2.40	0.47	0.00	3.95	208
SULFUR DIOXIDE	7.14	2.37	5.93	2.99	2.50	3.95	279
4-AMINOPYRIDINE	10.53	0.00	0.00	0.00	0.00	3.88	21
CUBE EXTRACTS	5.26	0.00	0.00	0.00	0.00	3.77	21
METHYL BROMIDE	3.91	3.80	3.97	2.53	2.93	3.43	8,225
THIABENDAZOLE	3.56	8.58	2.79	2.13	0.85	3.30	1,851

Table 5a. The total number of pounds of all pesticide active ingredients used in each county in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 1a are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data is sorted by percent change.

County	All data Total lbs AI	Criterion 1a Total lbs AI	Change
TULARE	15,739,984	15,126,456	4.06
SAN DIEGO	996,653	964,234	3.36
IMPERIAL	8,003,543	7,858,570	1.84
SAN JOAQUIN	11,447,637	11,397,731	0.44
SAN BERNADINO	442,140	440,666	0.33
ORANGE	936,416	934,421	0.21
YOLO	3,025,381	3,019,343	0.20
VENTURA	5,448,998	5,445,094	0.07
MONTEREY	12,478,083	12,471,576	0.05
TEHAMA	859,370	859,027	0.04
PLACER	223,527	223,490	0.02
SUTTER	3,406,431	3,406,161	0.01
SONOMA	3,882,171	3,881,877	0.01
SANTA BARBARA	3,202,009	3,201,813	0.01
FRESNO	38,154,227	38,153,138	0.00
ALAMEDA	138,458	138,458	0.00
AMADOR	126,682	126,682	0.00
BUTTE	3,373,754	3,373,754	0.00
CALAVERAS	33,136	33,136	0.00
COLUSA	2,818,773	2,818,773	0.00
CONTRA COSTA	312,692	312,692	0.00
DEL NORTE	218,006	218,006	0.00
EL DORADO	83,926	83,926	0.00
GLENN	2,219,654	2,219,654	0.00
HUMBOLDT	56,011	56,011	0.00
INYO	7,336	7,336	0.00
KERN	23,361,687	23,361,687	0.00
KINGS	5,247,248	5,247,248	0.00
LAKE	941,624	941,624	0.00
LASSEN	120,684	120,684	0.00
LOS ANGELES	111,850	111,850	0.00
MADERA	9,204,383	9,204,383	0.00
MARIN	8,542	8,542	0.00
MARIPOSA	3,906	3,906	0.00
MENDOCINO	1,888,216	1,888,216	0.00
MERCED	7,034,047	7,034,047	0.00
MODOC	139,011	139,011	0.00
MONO	11,511	11,511	0.00
NAPA	2,824,865	2,824,865	0.00
NEVADA	13,884	13,884	0.00
RIVERSIDE	4,234,696	4,234,696	0.00
SACRAMENTO	2,272,006	2,272,006	0.00
SAN BENITO	593,440	593,440	0.00
SAN FRANCISCO	19	19	0.00
SAN LOUIS OBISPO	1,600,508	1,600,508	0.00
SAN MATEO	96,514	96,514	0.00
SANTA CLARA	266,236	266,236	0.00
SANTA CRUZ	1,658,271	1,658,271	0.00
SHASTA	316,946	316,946	0.00
SISKIYOU	428,211	428,211	0.00

Table 5a. The total number of pounds of all pesticide active ingredients used in each county in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 1a are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data is sorted by percent change.

County	All data Total lbs AI	Criterion 1a Total lbs AI	Change
SOLANO	1,590,363	1,590,363	0.00
STANISLAUS	5,044,375	5,044,375	0.00
TRINITY	580	580	0.00
TUOLUMNE	5,427	5,427	0.00
YUBA	1,711,224	1,711,224	0.00
TOTAL	188,365,271	187,502,299	0.46

Table 5b. The total number of pounds of all pesticide active ingredients used in each county in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 2b are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data is sorted by percent change.

County	All data Total lbs AI	Criterion 2b Total lbs AI	Change
MARIPOSA	3,906	3,506	11.41
TULARE	15,739,984	15,074,899	4.41
MONTEREY	12,478,083	12,036,394	3.67
SAN MATEO	96,514	94,174	2.48
MODOC	139,011	135,753	2.40
YOLO	3,025,381	2,959,273	2.23
LAKE	941,624	923,145	2.00
ALAMEDA	138,458	135,795	1.96
CALAVERAS	33,136	32,505	1.94
SANTA CLARA	266,236	261,173	1.94
BUTTE	3,373,754	3,310,959	1.90
SUTTER	3,406,431	3,354,429	1.55
SANTA BARBARA	3,202,009	3,155,686	1.47
COLUSA	2,818,773	2,781,942	1.32
LOS ANGELES	111,850	110,445	1.27
CONTRA COSTA	312,692	308,880	1.23
SAN JOAQUIN	11,447,637	11,310,848	1.21
STANISLAUS	5,044,375	4,984,102	1.21
SAN BERNADINO	442,140	437,251	1.12
RIVERSIDE	4,234,696	4,202,934	0.76
SAN DIEGO	996,653	990,883	0.58
ORANGE	936,416	931,071	0.57
EL DORADO	83,926	83,451	0.57
SAN BENITO	593,440	590,262	0.54
IMPERIAL	8,003,543	7,967,000	0.46
SACRAMENTO	2,272,006	2,262,802	0.41
KERN	23,361,687	23,268,723	0.40
MADERA	9,204,383	9,167,971	0.40
FRESNO	38,154,227	38,015,006	0.37
PLACER	223,527	222,722	0.36
MERCED	7,034,047	7,010,371	0.34
GLENN	2,219,654	2,212,360	0.33
VENTURA	5,448,998	5,436,481	0.23
SONOMA	3,882,171	3,874,505	0.20
HUMBOLDT	56,011	55,918	0.17
MENDOCINO	1,888,216	1,885,208	0.16
SISKIYOU	428,211	427,559	0.15
KINGS	5,247,248	5,240,109	0.14
YUBA	1,711,224	1,709,152	0.12
SANTA CRUZ	1,658,271	1,656,595	0.10
SOLANO	1,590,363	1,589,030	0.08
AMADOR	126,682	126,599	0.07
TEHAMA	859,370	858,993	0.04
SAN LOUIS OBISPO	1,600,508	1,600,041	0.03
LASSEN	120,684	120,674	0.01
NAPA	2,824,865	2,824,791	0.00
MARIN	8,542	8,542	0.00
DEL NORTE	218,006	218,006	0.00
INYO	7,336	7,336	0.00
MONO	11,511	11,511	0.00
NEVADA	13,884	13,884	0.00

Table 5b. The total number of pounds of all pesticide active ingredients used in each county in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 2b are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data is sorted by percent change.

County	All data Total lbs AI	Criterion 2b Total lbs AI	Change
SAN FRANCISCO	19	19	0.00
SHASTA	316,946	316,946	0.00
TRINITY	580	580	0.00
TUOLUMNE	5,427	5,427	0.00
TOTAL	188,365,271	186,324,619	1.10

Table 5c. The total number of pounds of all pesticide active ingredients used in each county in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 4d are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data is sorted by percent change.

County	All data Total lbs AI	Criterion 4d Total lbs AI	Change
DEL NORTE	218,006	191,736	13.70
MARIPOSA	3,906	3,506	11.41
CONTRA COSTA	312,692	294,437	6.20
MADERA	9,204,383	8,745,105	5.25
TULARE	15,739,984	14,990,358	5.00
SAN MATEO	96,514	92,896	3.89
MONTEREY	12,478,083	12,011,219	3.89
COLUSA	2,818,773	2,715,607	3.80
BUTTE	3,373,754	3,277,072	2.95
SUTTER	3,406,431	3,326,129	2.41
MODOC	139,011	135,753	2.40
YOLO	3,025,381	2,956,375	2.33
LAKE	941,624	921,204	2.22
ALAMEDA	138,458	135,683	2.05
SANTA CLARA	266,236	260,942	2.03
CALAVERAS	33,136	32,488	2.00
SANTA BARBARA	3,202,009	3,141,177	1.94
SAN DIEGO	996,653	978,171	1.89
TEHAMA	859,370	843,486	1.88
ORANGE	936,416	919,508	1.84
SAN JOAQUIN	11,447,637	11,251,416	1.74
LOS ANGELES	111,850	110,100	1.59
NAPA	2,824,865	2,782,033	1.54
SAN LOUIS OBISPO	1,600,508	1,577,250	1.47
SONOMA	3,882,171	3,826,321	1.46
SAN BERNADINO	442,140	436,378	1.32
RIVERSIDE	4,234,696	4,179,667	1.32
YUBA	1,711,224	1,689,951	1.26
SAN BENITO	593,440	587,077	1.08
KERN	23,361,687	23,137,656	0.97
FRESNO	38,154,227	37,789,450	0.97
PLACER	223,527	222,150	0.62
IMPERIAL	8,003,543	7,954,644	0.61
GLENN	2,219,654	2,206,420	0.60
EL DORADO	83,926	83,444	0.58
SACRAMENTO	2,272,006	2,260,132	0.53
SHASTA	316,946	315,290	0.53
VENTURA	5,448,998	5,424,992	0.44
STANISLAUS	5,044,375	5,026,342	0.36
MERCED	7,034,047	7,011,006	0.33
SOLANO	1,590,363	1,585,484	0.31
SANTA CRUZ	1,658,271	1,653,703	0.28
KINGS	5,247,248	5,233,987	0.25
AMADOR	126,682	126,442	0.19
MENDOCINO	1,888,216	1,884,691	0.19
HUMBOLDT	56,011	55,917	0.17
NEVADA	13,884	13,861	0.16
SISKIYOU	428,211	427,559	0.15
LASSEN	120,684	120,674	0.01
MARIN	8,542	8,542	0.00

Table 5c. The total number of pounds of all pesticide active ingredients used in each county in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 4d are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data is sorted by percent change.

County	All data Total lbs AI	Criterion 4d Total lbs AI	Change
INYO	7,336	7,336	0.00
MONO	11,511	11,511	0.00
SAN FRANCISCO	19	19	0.00
TRINITY	580	580	0.00
TUOLUMNE	5,427	5,427	0.00
TOTAL	188,365,271	184,980,305	1.83

Table 5d. The total number of pounds of all pesticide active ingredients used in each county in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criteria 1a, 2b, or 4d are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data is sorted by percent change.

	All data	Criteria 1a, 2b, 4d	
County	**Total lbs AI**	**Total lbs AI**	**Change**
DEL NORTE	218,006	191,736	13.70
MARIPOSA	3,906	3,506	11.41
CONTRA COSTA	312,692	294,437	6.20
SAN DIEGO	996,653	945,709	5.39
MADERA	9,204,383	8,745,105	5.25
TULARE	15,739,984	14,971,216	5.13
BUTTE	3,373,754	3,225,471	4.60
MONTEREY	12,478,083	12,004,648	3.94
SAN MATEO	96,514	92,887	3.90
COLUSA	2,818,773	2,715,607	3.80
SAN JOAQUIN	11,447,637	11,137,821	2.78
SUTTER	3,406,431	3,317,090	2.69
IMPERIAL	8,003,543	7,809,670	2.48
MODOC	139,011	135,753	2.40
YOLO	3,025,381	2,956,375	2.33
LAKE	941,624	921,204	2.22
ALAMEDA	138,458	135,537	2.16
SANTA CLARA	266,236	260,939	2.03
CALAVERAS	33,136	32,488	2.00
SANTA BARBARA	3,202,009	3,139,617	1.99
TEHAMA	859,370	843,143	1.92
ORANGE	936,416	919,057	1.89
SAN BERNADINO	442,140	434,899	1.66
LOS ANGELES	111,850	110,100	1.59
NAPA	2,824,865	2,782,033	1.54
SAN LOUIS OBISPO	1,600,508	1,577,247	1.47
SONOMA	3,882,171	3,825,989	1.47
STANISLAUS	5,044,375	4,971,574	1.46
YUBA	1,711,224	1,687,903	1.38
RIVERSIDE	4,234,696	4,179,541	1.32
SAN BENITO	593,440	587,077	1.08
FRESNO	38,154,227	37,787,711	0.97
KERN	23,361,687	23,137,656	0.97
PLACER	223,527	222,150	0.62
GLENN	2,219,654	2,206,420	0.60
EL DORADO	83,926	83,444	0.58
MERCED	7,034,047	6,995,038	0.56
SACRAMENTO	2,272,006	2,260,132	0.53
SHASTA	316,946	315,290	0.53
VENTURA	5,448,998	5,421,088	0.51
SOLANO	1,590,363	1,585,484	0.31
SANTA CRUZ	1,658,271	1,653,703	0.28
KINGS	5,247,248	5,233,987	0.25
AMADOR	126,682	126,442	0.19
MENDOCINO	1,888,216	1,884,691	0.19
HUMBOLDT	56,011	55,917	0.17
NEVADA	13,884	13,861	0.16
SISKIYOU	428,211	427,559	0.15
LASSEN	120,684	120,674	0.01

Table 5d. The total number of pounds of all pesticide active ingredients used in each county in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criteria 1a, 2b, or 4d are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data is sorted by percent change.

County	All data Total lbs AI	Criteria 1a, 2b, 4d Total lbs AI	Change
MARIN	8,542	8,542	0.00
INYO	7,336	7,336	0.00
MONO	11,511	11,511	0.00
SAN FRANCISCO	19	19	0.00
TRINITY	580	580	0.00
TUOLUMNE	5,427	5,427	0.00
TOTAL	188,365,271	184,520,043	2.08

Table 6a. The total number of pounds of different pesticide active ingredients in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 1a are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data are sorted by percent change and only chemicals in which outliers were found are shown.

Active Ingredient	All data Total lbs AI	Criterion 1a Total lbs AI	Change
CARBARYL	1,429,209	832,698	71.64
PEBULATE	253,283	247,245	2.44
XYLENE RANGE AROMATIC SOLVENT	38,404	38,008	1.04
METAM-SODIUM	15,221,984	15,077,011	0.96
METHYL BROMIDE	16,673,970	16,577,249	0.58
ACROLEIN	87,023	86,680	0.40
MALATHION	676,819	674,170	0.39
PETROLEUM DISTILLATES, REFINED	39,483	39,391	0.23
DIAZINON	925,076	923,173	0.21
POTASH SOAP	293,272	292,676	0.20
CHLOROPICRIN	2,807,525	2,803,186	0.15
MINERAL OIL	3,341,236	3,339,762	0.04
GLYPHOSATE, ISOPROPYLAMINE SALT	2,740,463	2,739,260	0.04
PETROLEUM OIL, UNCLASSIFIED	18,986,513	18,980,815	0.03
2,4-D, DIMETHYLAMINE SALT	400,254	400,218	0.01

Table 6b. The total number of pounds of different pesticide active ingredients in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 2b are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data are sorted by percent change, only the highest 50 chemicals are shown, and only chemicals with more than 10 records.

	All data	Criterion 2b	
Active Ingredient	**Total lbs AI**	**Total lbs AI**	**Change**
AGROBACTERIUM RADIOBACTER	208	3	6,867.41
DIDECYL DIMETHYL AMMONIUM CHLORIDE	1,119	49	2,193.46
DIOCTYL DIMETHYL AMMONIUM CHLORIDE	1,119	49	2,193.46
OCTYL DECYL DIMETHYL AMMONIUM CHLORIDE	2,237	98	2,193.46
ALKYL(50%C14,40%C12,10%C16)DIMETHYLBENZYL AMMONIUM CH	3,011	159	1,798.18
8-DODECENE-1-OL, OTHER RELATED	34	2	1,339.53
CHLORSULFURON	5,067	379	1,235.90
(Z,E) 7,11 HEXADECADIEN-1-01 ACETATE	392	29	1,230.71
Z-8-DODECENOL	36	6	520.95
E-8-DODECENYL ACETATE	238	39	514.35
Z-8-DODECENYL ACETATE	4,100	673	509.53
PETROLEUM DISTILLATES	8,135	1,553	423.89
6-METHYL-1,3-DITHIOLO(4,5-B)QUINOXALIN-2-ONE	2,934	588	399.22
KINOPRENE	4,501	1,108	306.27
MYRISTYL ALCOHOL	366	117	211.40
LAURYL ALCOHOL	1,784	580	207.85
E,E-8,10-DODECADIEN-1-OL	3,236	1,067	203.15
TERRAZOLE	187	72	160.61
NICOSULFURON	3,408	1,428	138.63
2,4-D, ISOPROPYL ESTER	11,479	5,542	107.14
DIFLUBENZURON	13,841	7,282	90.06
CARBARYL	1,429,209	827,584	72.70
BENSULFURON METHYL	45,122	26,268	71.78
GIBBERELLINS	41,650	25,243	64.99
PACLOBUTRAZOL	35	24	45.28
GLYPHOSATE, MONOAMMONIUM SALT	6,786	4,723	43.69
TAU FLUVALINATE	4,824	3,537	36.39
ALUMINUM PHOSPHIDE	40,195	29,887	34.49
DIENOCHLOR	9,443	7,056	33.82
PROPYLENE OXIDE	105,470	81,173	29.93
CHLOROPHACINONE	5	4	29.76
ALKYL(68%C12, 32%C14)DIMETHYL ETHYLBENZYL AMMONIUM CHL	350	274	28.03
MAGNESIUM PHOSPHIDE	833	651	27.88
CHLORPROPHAM	4,103	3,230	27.01
CYFLUTHRIN	17,579	14,089	24.77
(E)-5-DECENYL ACETATE	71	58	22.90
(E)-5-DECENOL	15	12	22.90
CHLORMEQUAT CHLORIDE	1,158	947	22.29
NAA	4	3	19.41
IBA	10	9	17.76
ALKYL(60%C14,30%C16,5%C12,5%C18)DIMETHYL BENZYL AMMONI	583	506	15.15
NAPROPAMIDE	220,718	192,154	14.87
IMIDACLOPRID	63,169	55,103	14.64
THIABENDAZOLE	18,422	16,220	13.58
CYPERMETHRIN	37,788	33,322	13.40
CHLORINE	3,163,696	2,795,882	13.16
MCPA, DIMETHYLAMINE SALT	288,495	255,838	12.76
TRIADIMEFON	18,238	16,261	12.16
ARSENIC ACID	38,072	34,037	11.86
AVERMECTIN	7,881	7,078	11.34

Table 6c. The total number of pounds of different pesticide active ingredients in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 4d are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data are sorted by percent change, only the highest 50 chemicals are shown, and only chemicals with more than 10 records.

Active Ingredient	All data Total lbs AI	Criterion 4d Total lbs AI	Change
AGROBACTERIUM RADIOBACTER	208	3	6,867.41
DIDECYL DIMETHYL AMMONIUM CHLORIDE	1,119	49	2,200.96
DIOCTYL DIMETHYL AMMONIUM CHLORIDE	1,119	49	2,200.96
OCTYL DECYL DIMETHYL AMMONIUM CHLORIDE	2,237	97	2,200.96
ALKYL(50%C14,40%C12,10%C16)DIMETHYLBENZYL AMMONIUM	3,011	158	1,803.27
8-DODECENE-1-OL, OTHER RELATED	34	2	1,339.53
CHLORSULFURON	5,067	379	1,235.90
(Z,E) 7,11 HEXADECADIEN-1-01 ACETATE	392	29	1,230.71
Z-8-DODECENOL	36	6	520.95
E-8-DODECENYL ACETATE	238	39	514.35
Z-8-DODECENYL ACETATE	4,100	673	509.53
PETROLEUM DISTILLATES	8,135	1,552	424.13
6-METHYL-1,3-DITHIOLO(4,5-B)QUINOXALIN-2-ONE	2,934	581	404.93
KINOPRENE	4,501	1,067	321.79
MYRISTYL ALCOHOL	366	117	211.40
LAURYL ALCOHOL	1,784	580	207.85
E,E-8,10-DODECADIEN-1-OL	3,236	1,067	203.21
TERRAZOLE	187	72	160.61
NICOSULFURON	3,408	1,388	145.59
2,4-D, ISOPROPYL ESTER	11,479	4,922	133.22
GIBBERELLINS	41,650	21,092	97.47
DIFLUBENZURON	13,841	7,134	94.01
BENSULFURON METHYL	45,122	23,811	89.50
CARBARYL	1,429,209	811,729	76.07
PACLOBUTRAZOL	35	24	48.55
GLYPHOSATE, MONOAMMONIUM SALT	6,786	4,723	43.69
TAU FLUVALINATE	4,824	3,394	42.13
COPPER ETHANOLAMINE COMPLEXES, MIXED	1,420	1,008	40.83
ALUMINUM PHOSPHIDE	40,195	29,134	37.97
DIENOCHLOR	9,443	6,900	36.85
CHLOROPHACINONE	5	4	30.95
PROPYLENE OXIDE	105,470	81,173	29.93
CHLORMEQUAT CHLORIDE	1,158	897	29.17
ALKYL(68%C12, 32%C14)DIMETHYL ETHYLBENZYL AMMONIUM	350	273	28.13
MAGNESIUM PHOSPHIDE	833	651	27.88
CHLORPROPHAM	4,103	3,230	27.01

Table 6c. The total number of pounds of different pesticide active ingredients in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criterion 4d are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data are sorted by percent change, only the highest 50 chemicals are shown, and only chemicals with more than 10 records.

	All data	Criterion 4d	
Active Ingredient	**Total lbs AI**	**Total lbs AI**	**Change**
CYFLUTHRIN	17,579	13,948	26.04
IMAZALIL	13,553	10,867	24.71
FENARIMOL	22,541	18,127	24.35
(E)-5-DECENYL ACETATE	71	58	22.90
(E)-5-DECENOL	15	12	22.90
SAWDUST	375	306	22.60
PHOSPHOROUS	346	282	22.60
CARBON	1,847	1,507	22.59
SODIUM NITRATE	4,619	3,768	22.59
THIABENDAZOLE	18,422	15,078	22.17
ORTHO-PHENYLPHENOL, SODIUM SALT	32,907	27,009	21.84
NAA	4	3	19.41
MYCLOBUTANIL	100,945	85,620	17.90
IBA	10	9	17.79

Table 6d. The total number of pounds of different pesticide active ingredients in California in 1995. Data includes only records for which unit treated is greater that 0 but does not include adjuvants. The first column gives the total pounds used calculated from all the data currently in the PUR tables. The other columns give the total pounds and percentage change when all records that meet criteria 1a, 2b, or 4d are removed from the database. The percent change is calculated by (lbs all data - lbs without outliers) / lbs without outliers X 100. Data are sorted by percent change, only the highest 50 chemicals are shown, and only chemicals with more than 10 records.

	All data	Criteria 1a, 2b, 4d	
Active Ingredient	**Total lbs AI**	**Total lbs AI**	**Change**
AGROBACTERIUM RADIOBACTER	208	3	6,867.41
DIDECYL DIMETHYL AMMONIUM CHLORIDE	1,119	49	2,200.96
DIOCTYL DIMETHYL AMMONIUM CHLORIDE	1,119	49	2,200.96
OCTYL DECYL DIMETHYL AMMONIUM CHLORIDE	2,237	97	2,200.96
ALKYL(50%C14,40%C12,10%C16)DIMETHYLBENZYL AMMONIUM CHL(	3,011	158	1,803.27
8-DODECENE-1-OL, OTHER RELATED	34	2	1,339.53
CHLORSULFURON	5,067	379	1,235.90
(Z,E) 7,11 HEXADECADIEN-1-01 ACETATE	392	29	1,230.71
Z-8-DODECENOL	36	6	520.95
E-8-DODECENYL ACETATE	238	39	514.35
Z-8-DODECENYL ACETATE	4,100	673	509.53
PETROLEUM DISTILLATES	8,135	1,552	424.13
6-METHYL-1,3-DITHIOLO(4,5-B)QUINOXALIN-2-ONE	2,934	581	404.93
KINOPRENE	4,501	1,067	321.79
MYRISTYL ALCOHOL	366	117	211.40
LAURYL ALCOHOL	1,784	580	207.85
E,E-8,10-DODECADIEN-1-OL	3,236	1,067	203.21
TERRAZOLE	187	72	160.61
NICOSULFURON	3,408	1,388	145.59
2,4-D, ISOPROPYL ESTER	11,479	4,922	133.22
GIBBERELLINS	41,650	21,092	97.47
DIFLUBENZURON	13,841	7,134	94.01
BENSULFURON METHYL	45,122	23,811	89.50
CARBARYL	1,429,209	811,729	76.07
PACLOBUTRAZOL	35	24	48.62
GLYPHOSATE, MONOAMMONIUM SALT	6,786	4,723	43.69
TAU FLUVALINATE	4,824	3,394	42.13
COPPER ETHANOLAMINE COMPLEXES, MIXED	1,420	1,008	40.83
ALUMINUM PHOSPHIDE	40,195	29,070	38.27
DIENOCHLOR	9,443	6,900	36.85
CHLOROPHACINONE	5	4	30.95
PROPYLENE OXIDE	105,470	81,173	29.93
ALKYL(68%C12, 32%C14)DIMETHYL ETHYLBENZYL AMMONIUM CHLC	350	271	29.31
CHLORMEQUAT CHLORIDE	1,158	897	29.17
MAGNESIUM PHOSPHIDE	833	651	27.88
CHLORPROPHAM	4,103	3,230	27.01
CYFLUTHRIN	17,579	13,948	26.04
IMAZALIL	13,553	10,867	24.71
FENARIMOL	22,541	18,127	24.35
(E)-5-DECENYL ACETATE	71	58	22.90
(E)-5-DECENOL	15	12	22.90
SAWDUST	375	306	22.60
PHOSPHOROUS	346	282	22.60
CARBON	1,847	1,507	22.59
SODIUM NITRATE	4,619	3,768	22.59
THIABENDAZOLE	18,422	15,078	22.17
ORTHO-PHENYLPHENOL, SODIUM SALT	32,907	27,009	21.84
NAA	4	3	19.41
MYCLOBUTANIL	100,945	85,620	17.90
IBA	10	9	17.79

Appendix I: Further explanation of the outlier criteria

A better understanding of the different criteria and their respective advantages and disadvantages can be gained by looking at a large number of frequency distributions of use rates with the outlier limits of each criterion superimposed (Fig. 1). Figure 1 shows a number of distributions for different active ingredients that were chosen because they had an unusually high percentage of outliers. Distributions are shown for only a few representative products that had these active ingredients. Each distribution shows the number of records for different use rates for a particular use type (that is, for a particular pesticide product, site, unit treated, and record type). The x-axis gives the rate of use divided by the median value of all rates of the same use type. These graphs also have superimposed the highest limit values for criteria 2, 3, and 4 (that is, criteria 2b, 3b, and 4d). For criterion 1, the lower limit value (1a) was used in Fig. 1 because it turned out to be extremely high in most cases.

<u>Criterion 1: Pounds per acre of active ingredient is larger than 200 or 400 (non-fumigants), or 1000 or 2000 (fumigants).</u>

The simplest criterion is to flag values that are larger than some predetermined extreme value, but the rates of use for different active ingredients can vary considerably from only a few ounces per acre to hundreds of pounds per acre. To reduce this problem all pesticides were divided into two groups. One group included pesticides that are often used at rates of hundreds of pounds per acre, mostly fumigants (methyl bromide, metam-sodium, chloropicrin, dazomet, boric acid, and 1,3 dichloropropene). The other group included all other pesticides, most of which are seldom applied at rates over 100 pounds per acre.

One difficulty in implementing this criterion is that the pounds of pesticide recorded in the PUR is for a pesticide product which includes other constituents besides the active ingredient. Also, some products contain more than one active ingredient. To deal with this problem, the prod_chem database was queried to get a list of all the active ingredients in the product and the percentage by weight of these active ingredients. The pounds of each active ingredient was then calculated. Since the goal is to flag records in which at least one of the active ingredients in the product was higher than the extreme value, only the pounds of active ingredient with the highest percentage was used to flag the record or not. Since there are two groups of pesticides with different extreme values, the highest uses for both groups were identified.

This criterion obviously is not affected by the type of distribution so it can be applied to any type of distribution (Table 1). In particular it can be applied even when there is only one record of a particular use type. However, if the typical use rates of a pesticide are high, criterion 1 can make type I errors. Similarly, if typical use rates are low, criterion 1 can make type II errors. Finally, because criterion 1 only applies to records with units in acres, it will miss outliers in any record measured with any other unit and so make many type II errors.

The limit value for criterion 1a (Fig. 1) varies tremendously for different pesticides, from a normalized rate (that is, the rate divided by the median) of 0.6 to 124,069,480. The reason for this variance is that the usual rate of use of different pesticides varies tremendously. This fact was partly accounted for by dividing pesticides into two groups (fumigants and non-fumigants), but even within each of these groups there is a huge variation in use rates. This variation, which is independent of the nature of the distribution of the rates of use, is the major disadvantage of criterion 1.

In most cases shown in Fig. 1 the criterion 1a limit is too large relative to typical use rates (and, of course, criterion 1b is even larger) and thus type II errors are common. It will also create type II errors for outliers that occur for any use on units that were not in acres (Fig. 1m, w, γ, δ). However, a few type I errors probably also occur (Fig. 1i, 1j, 1k, 1l, 1u). Most of these cases are the fumigants, which suggest that the criterion 1 limit value for fumigants was set too low. The apparently unusually low limit value for criterion 1a in Fig 1u will be explained when criterion 3 is discussed.

<u>Criterion 2: Pounds per unit treated of a product is larger than 25 or 50 times the median.</u>

Criterion 2 is also fairly simple but it improves on criterion 1 because it takes into account the typical use rate of the different pesticides. Since the label rates of each pesticide are not available in DPR's databases, a reasonable rate of use for each product on each site was estimated by calculating the median pounds of a pesticide product (not just pounds of active ingredient) per unit treated of all records of that use type. The unit treated could be acres, square feet, cubic feet, etc. In order to minimize the variation in use rates, the medians were calculated for a group of records with the same use type (that is, the same pesticide product, on the same site, for the same unit treated, and the same record type). There are two general record types that differ in how pesticide records are processed. One record type refers to structural and rights of way uses (in the PUR these have values of 'C', '2', and 'G' in the record_id field) and the other are production agricultural uses (record_ids 'A', 'B', '1', '4', 'E', and 'F'). Presumably, the rates of all applications for each use type should be similar to one another.

Thus, criterion 2 is an improvement over criterion 1 because it can be useful whether the typical uses are either high or low (Table 1). It also is an improvement over criterion 1 because it can be used for records treated on any unit, not just acres.

However, it has some disadvantages relative to criterion 1. First, there must be other records of the same use type so that a comparison can be made. Obviously, if there is only one record, no comparison can be made at all. And if there are only a few records, it may be that all are outliers and these would not be picked up by criterion 2 (type II error). Also, if the usual range of uses is very small, type II errors can occur. However, type I errors can also occur if there is a broad range of use rates for some pesticide use. That is, it may be that rates of even 50 times the median value is a valid rate for some kinds of uses. A type I error could also occur if over half of the records were in error by being over 50 times too small. In this case, the few valid records would be flagged as outliers. However, there is no way to look at a distribution to find this kind of error.

The criterion 2b limit is usually not nearly as large as criterion 1 limits, but often it too seems too large, causing type II errors (Fig. 1a, 1d, 1m, 1o, 1q, 1t, 1v, 1δ). Type I errors seem to be fairly rare, only occurring when there is a very broad distribution of use rates (possibly Fig. 1n).

<u>Criterion 3: Pounds per unit of product is larger than the median + 10 × median deviation or median + 50 × median deviation.</u>

Criterion 3 is a further development of criterion 2 by adding consideration of the distribution of the use rate values. That is, it increases the outlier limits for broad distributions and decreases it for narrow distributions, thus improving the main disadvantages of criterion 2. For example, if all records of some pesticide use type are between, say 1 and 4 pounds (with a median of 2), but one use mistakenly recorded 40 pounds, then criterion 2 would fail to identify it as an outlier because it is less than 25 times the median. This problem can be remedied by calculating some measure of dispersion, such as the standard deviation. Because we are using medians and want to minimize the effects of extreme values, for the measure of dispersion here we will use the median deviation, which is the median of the absolute values of the differences of each record with the median. As in criterion 2, the median is calculated for each different use type.

For normal distributions or distributions that are close to being normal, criterion 3 works very well. It still, of course, retains the other advantages of criterion 2: it works well whether usual rates are high or low and it applies to records with any kind of unit treated.

However, criterion 3 has some significant disadvantages. It still retains one of the disadvantages of criterion 2 -- if there are very few records, type II errors may occur. In addition, it often leads to type I errors in two new situations. If more than half of the records of a use type have the same rate, then the limit value is 1 no matter what the other rates are. The reason for this is that the median deviation in this situation is zero.

The second situation that leads to type I errors is when there is a multimodal distribution. In some cases, the criterion 3 limit could occur between two modes, which would mean that all values in the higher mode would be incorrectly flagged. However, type I errors could also occur with a multimodal distribution if, for example, there were two nearly equally large modes far apart from each other. In this case, criteria 3 would place the outlier limit too high because both the median and median deviations would be large.

These different situations are illustrated in Figure 1. In all cases where the distribution appears to be at least somewhat normal, criterion 3b seems to work very well in identifying outliers, that is, does not make either type I or type II errors (Figs. 1a, 1d, 1v).

However, several graphs illustrate the problem when more than half of the records have the same or nearly the same rate (Figs. 1b, 1f, 1g, 1h, 1i, 1j, 1l, 1m, 1p, 1s, 1ε). These uses are not close to being normally distributed. In all these cases, nearly half of the records will be flagged; most of which are perfectly reasonable values. Since this appears to a fairly common situation in the PUR, criterion 3 is seriously flawed.

The only immediately obvious example of a multimodal distribution is Fig. 1n. This distribution has three modes, one of which is approximately 100 times the median. Assuming that not all the 35 values in the highest mode are errors (which seems unlikely), then criterion 3b incorrectly places the outlier limit between the two higher modes. The outlier limit should probably be beyond the higher mode. The other type of multimodal distribution (with two equally large modes far apart from each other), which causes type I errors, is hidden by the way in which the graphs in Fig. 1 are presented. In Fig. 1, all of the values are divided by the median, putting the median at 1, so the two modes of a bimodal distribution would be clumped around 1. Examining the actual values of the distributions reveals that Fig. 1u is actually a bimodal distribution of this type. The five values in the histogram at 0.2 vary from 1.2 to 9.4, while the values in the other histograms vary from 400 to 1200. Note that the criterion 3b limit for this graph is higher than in any other graph. This distribution also explains why the limit for criterion 1a appears so small. It is indeed at a value of 200 pounds of active ingredient per acre but the median is even higher.

<u>Criterion 4: Pounds per unit of product is larger than a value generated using a neural network.</u>

Because criterion 3 failed in a couple of specialized (though common) situations, attempts were made to correct or improve it by various means. However, these procedures became more and more convoluted as new problems arose. So eventually, an entirely different procedure, using neural networks, was tried.

What is a neural network?
Even if a distribution is very unusual, people can look at it and make a judgment on what values are likely to be outliers. They do this by recognizing patterns and by accounting for a variety of circumstances and exceptions based on intuition and possibly from other experience in working with similar kinds of data. The pattern recognition ability of humans has been successfully imitated in many different situations by a computer programming technique known as neural networks. This technique was developed from attempts to build computers that operated on principles that were similar to the way human brains work—hence the name, which was derived from the biological neural network structure of brains.

A more accurate term for the type of neural network that is used in the outlier program is "artificial neural network" (ANN) since “neural networks” really refer to the biological structures in animal nervous systems. Analogous to the nervous systems neurons and axons, ANNs have nodes and connections (or weights). As in a biological neural network, an ANN consists of a set of nodes, each node having an output connection with many branches. Each branch of the output connection becomes an input connection to another node in the ANN set. The output value from a node is a function of all the input values that are connected to it.

An ANN is described mathematically by a vector function that relates a set (or vector) of input values (x_1, x_2, … x_n) to a vector of output values (y_1, y_2, … y_m). The heart of this function is another function that represents a single node:

$$n_j = \sigma\left(\sum_{i=1}^{N} w_{ij} n_i + w_j\right)$$

where N is the number of nodes, n_j is the output value of node j, n_i is the output value of node i (also one of the input values to node n_j), w_{ij} is the connection weight between input n_i and output node n_j, w_j is the threshold weight for node j, and σ is a function known as the activation function.

There are dozen of different possible ANN architectures (number of nodes, restrictions on number of connections or weights, the form of the activation function, etc). The outlier ANN uses what is known as a three-layer feedforward architecture with a sigmoid activation function given by

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

In a three layer ANN, the set of input values is known as the input layer and the set of output values is the output layer. The set of other nodes becomes the hidden layer. The input and output nodes are connected only to the hidden layer nodes. The number of input values, hidden nodes, and output values varies tremendously, depending on the particular use. The outlier ANN has 67 input values, 67 hidden nodes, 4 output values, and therefore $67 \times 67 + 67 \times 4 = 4757$ weights.

Training the artificial neural network

The most difficult part of creating an ANN is determining values for all the weight parameters. The process of determining these values is called “training” the neural network. Just as people learn from the time of birth by observing, remembering, and generalizing from many events in their experiences, an ANN also must learn by being presented with a large set of data. This training procedure involves presenting to the neural network program a set of data consisting of many examples of patterns (which are coded as input values) and the correct classification of each pattern (which are coded as output values). The program then adjusts the parameters in the neural network function until it produces the correct output values for each input as given in the training set.

There are many different algorithms for training neural network, but the most common, and the one used here, is known as the backpropagation method. Basically, this works by first creating a set of input vectors and corresponding correct output vectors (this set of input and output vectors is called the training set), which must be found or known in some way, depending on the application. How this training set was created for the outlier ANN will be described later.

To get the backpropagation process going, the weights are initially given random values. Then the process takes each input vector in the training set, uses the ANN function to calculate an output vector and compares this result with the correct output vector given in the training set. The difference between predicted and correct output vector becomes the error vector, which is used to adjust the weight parameters in such a way that the error vector becomes smaller. The parameters are adjusted for each input-output vector pair in the training set. This process is iterated many (usually thousands) of times through the training set until the error vector is less than some acceptable value for all the vectors in the training set.

If the vectors in the training set were representative of all the types of data of interest, then the ANN is said to be trained (that is, the weights are all given values). The ANN is now ready to be used to calculate acceptable (reasonably correct) output values for any set of input values.

Actually, developing a good neural network involves art as much as science. First of all, the training data must be sufficiently large and representative so it can correctly generalize all possible situations. If given a poor training set, the neural network will make poor predictions. Also, there are a large number of different types of training procedures and which one works best is subject to debate and probably at least partly depends on the type of application the neural network will be used for.

How was the PUR outlier neural network created?
The data used to train the neural network used in the PUR outlier program were generated from the pounds of pesticide product per unit treated for a selected set of pesticides and sites. Groups of pesticides and sites were chosen that included a wide range of types of distributions, including many unusual distributions. Two hundred frequency distributions were plotted and then these plots were examined by 12 scientists in DPR who marked values on each plot they thought were outliers. These scientists were asked to make two judgments for each distribution: values that they thought were obviously outliers and values they thought were suspicious outliers (the actual instructions that were given to them are reproduced in Appendix II).

For the neural network, the results from this survey were coded into numeric input and output values. The input values consisted of 67 statistical measures that characterized the distributions. The statistical measures included the number of values in the data set, mean, standard deviation, median, median deviation, skewness (a measure of asymmetry, in which one tail of the distribution is drawn out more than the other), and kurtosis (a measure of the peakedness of the distribution, in which there are more or less values near the mean relative to a normal distribution). The mean, standard deviation, skewness, and kurtosis were also calculated for a sequential series of 15 subsets of the data in which the highest values were removed. In particular, if there were less than 100 values in a set of rates, the highest rate value was removed and four statistics (mean, standard deviation, skewness, and kurtosis) were calculated for this smaller set. Then again, the next highest value was removed from the set and another set of statistics was calculated. These calculations were done for 15 subsets. If there were between 100 and 200 values, then each series removed the next two highest values, and so on.

The output values consisted of four values that were considered limit use rates. That is, any use rate above one of the limit values would be considered an outlier. These four limit values were determined in the following way: the highest limit (criterion 4d) was set for each distribution at a value that was just below all the values which all surveyees thought were obvious outliers; the next highest limit (criterion 4c) was set just below values which all surveyees thought were either obvious or probably outliers; the lowest limit (criterion 4a) was just below the values that only 1 to 3 surveyees thought were probable outliers; and the other limit (criterion 4b) was set between criteria limits 4a and 4c. Thus criterion 4 used four different limits to represent a range in confidence expressed by the surveyees as to the likelihood of being an outlier.

The 67 input and 4 output values for each of 180 survey distributions (90% of the 200 distributions given to the surveyees) were used to train the outlier ANN. After the neural network was trained, it was tested with the remaining 10% of the data set. The training procedure ran for thousands of iterations before it finally produced outputs that all were within 10% of the correct outputs. I tested the ANN by running the trained ANN with the input values from the remaining 20 distributions from the survey and compared the results with the correct output values. All the values agreed within 20%.

For the outlier ANN, a commercially available program, “BrainMaker”, produced by California Scientific Software, was used to train the ANN. The ANN function using the weight values produced by BrainMaker was programmed in Oracle ProC and run with the data from the PUR.

Evaluation of the ANN.

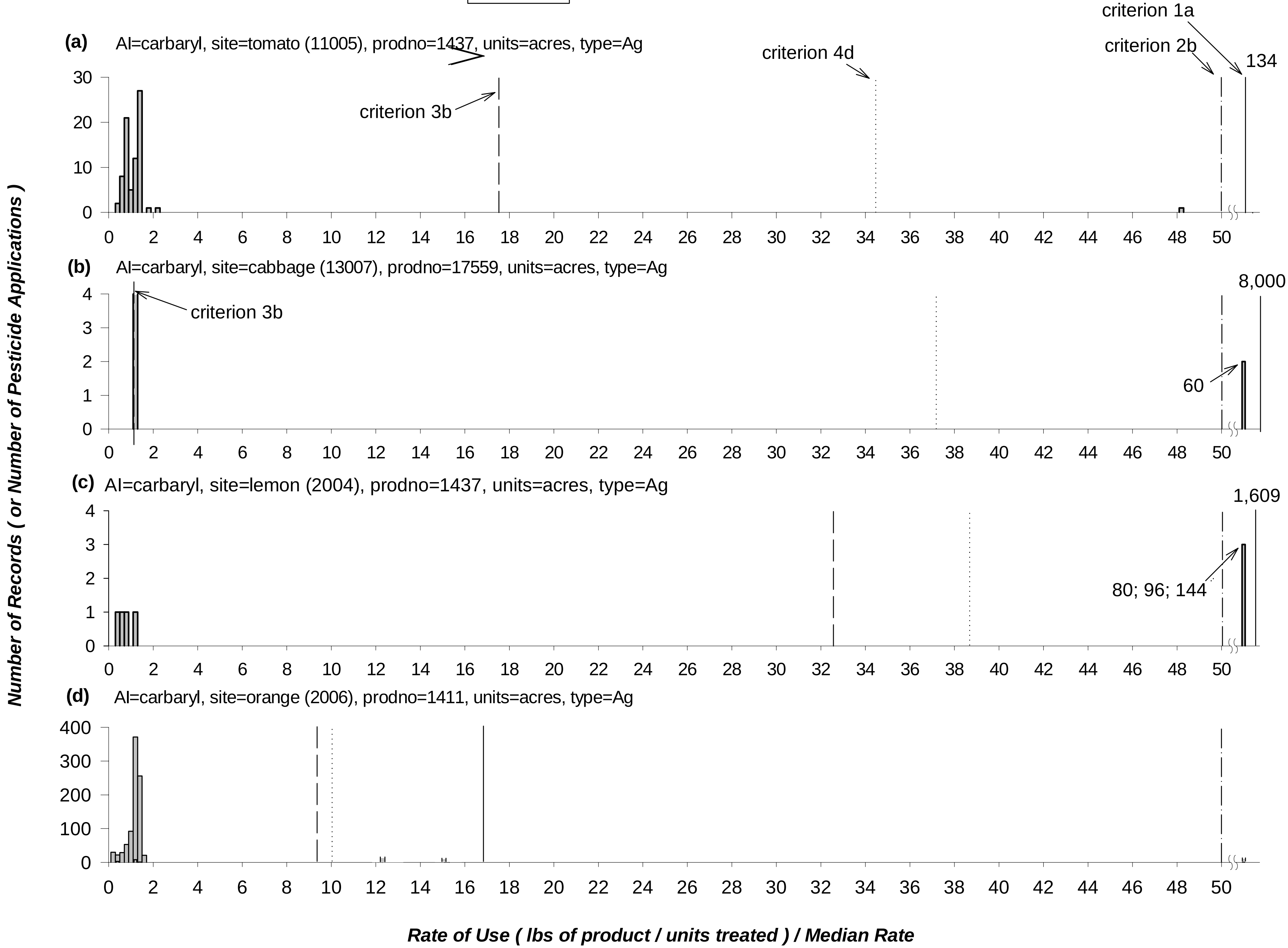
Neural networks worked very well in nearly all the types of situations where the other criteria failed (Table 1). Because, like humans, this procedure can recognize these unusual situations, it can adjust the limit values appropriately. The only situation where the procedure is likely to fail (producing type II errors) is when there are only a few records (where humans would fail too). That is, the neural network procedure, like criteria 2 and 3, must have a sufficient number of records to be able to know what are reasonable values for any use type. The only criterion that does not have this problem (at least for records with units in acres) is criterion 1. Note, too, that there are no situations where the neural network is likely to produce type I errors, which makes it a conservative criterion.

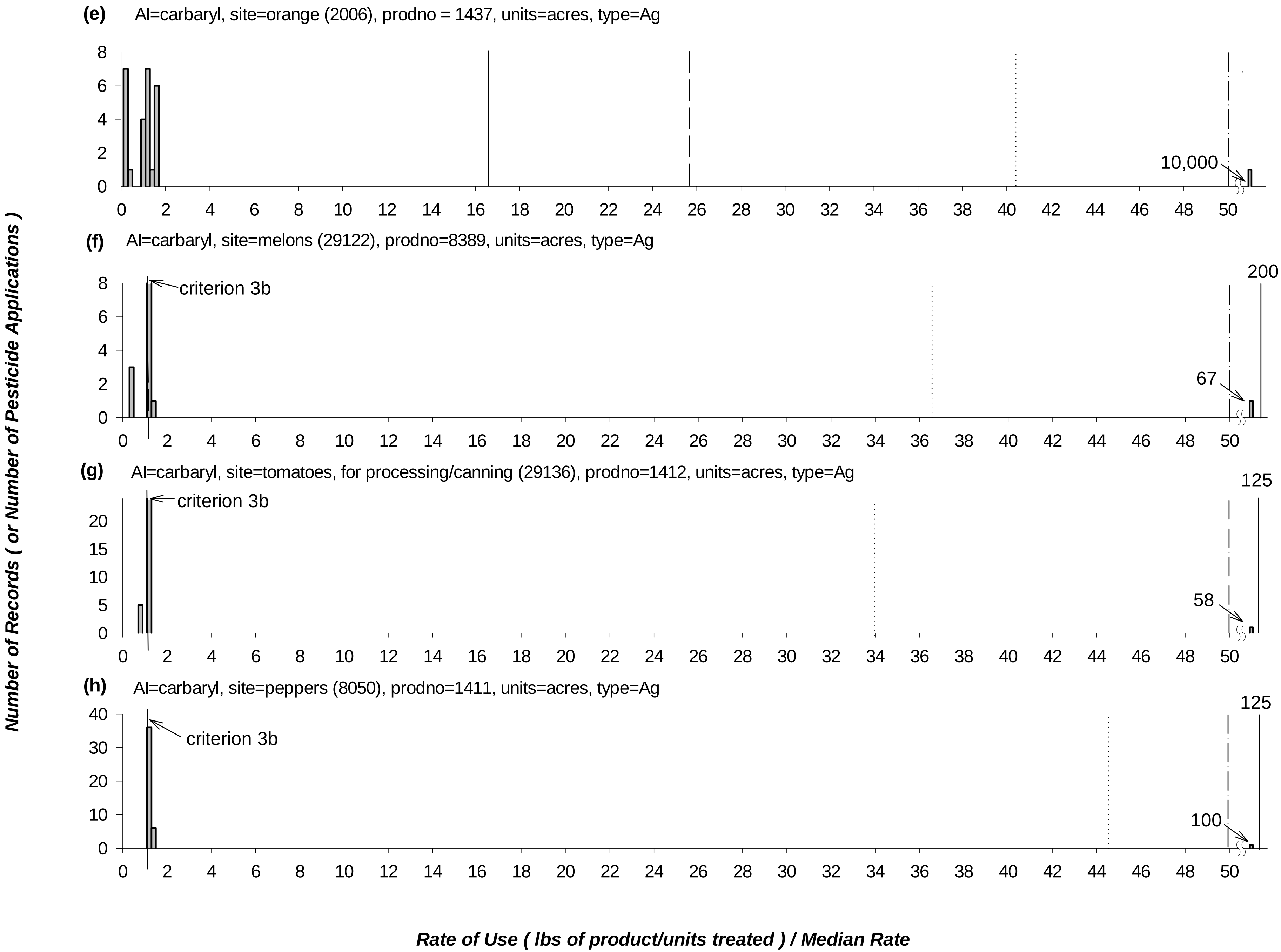
Because the neural network technique is a new and untried method for identifying outliers and because there is an element of art involved in training, the results of the neural network needs to be closely and extensively examined. The set of distributions in Fig. 1 illustrates that the neural network criterion works better, in most situations, than any of the other criteria.

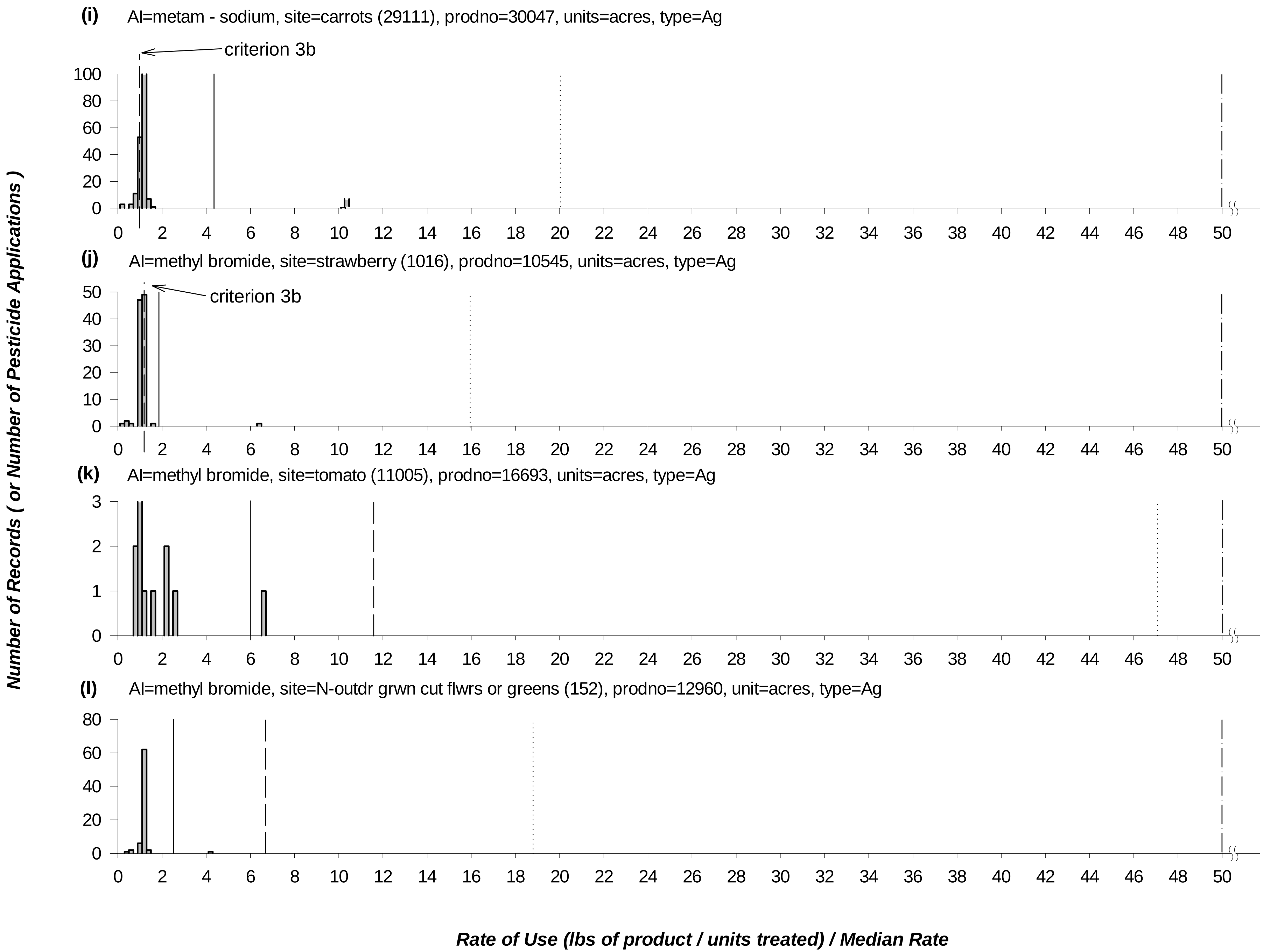
In most situations with normal distributions, criterion 4d limits are close to, but somewhat larger than, criterion 3b limits (Fig. 1). In situations where there are many records at one rate (cases where criteria 3b fails, Figs. 1b, 1f, 1g, 1h, 1i, 1j, 1l, 1m, 1p, 1s, 1ε), criteria 4d appears to give reasonable limit values. In the multimodal distribution mentioned above (Fig. 1n), criteria 4d appears to give a more reasonable outlier limit than criteria 3b. There are only two cases (Fig. 1t and 1u) where criterion 4d limit is less than criterion 3b limit (but still very close). Fig. 1t is an unusual distribution where there are a number of high rate values spread out over a very large range. It is difficult to say what is an outlier in this case. My opinion is that these are suspicious values, but I would not say they are definite outliers. If they are valid records, criteria 4d incorrectly includes them as outliers. Fig. 1u is the bimodal distribution discussed above, where it was concluded that the criterion 3b limit was probably too large. Thus, criterion 4d seems to be more reasonable in this case.

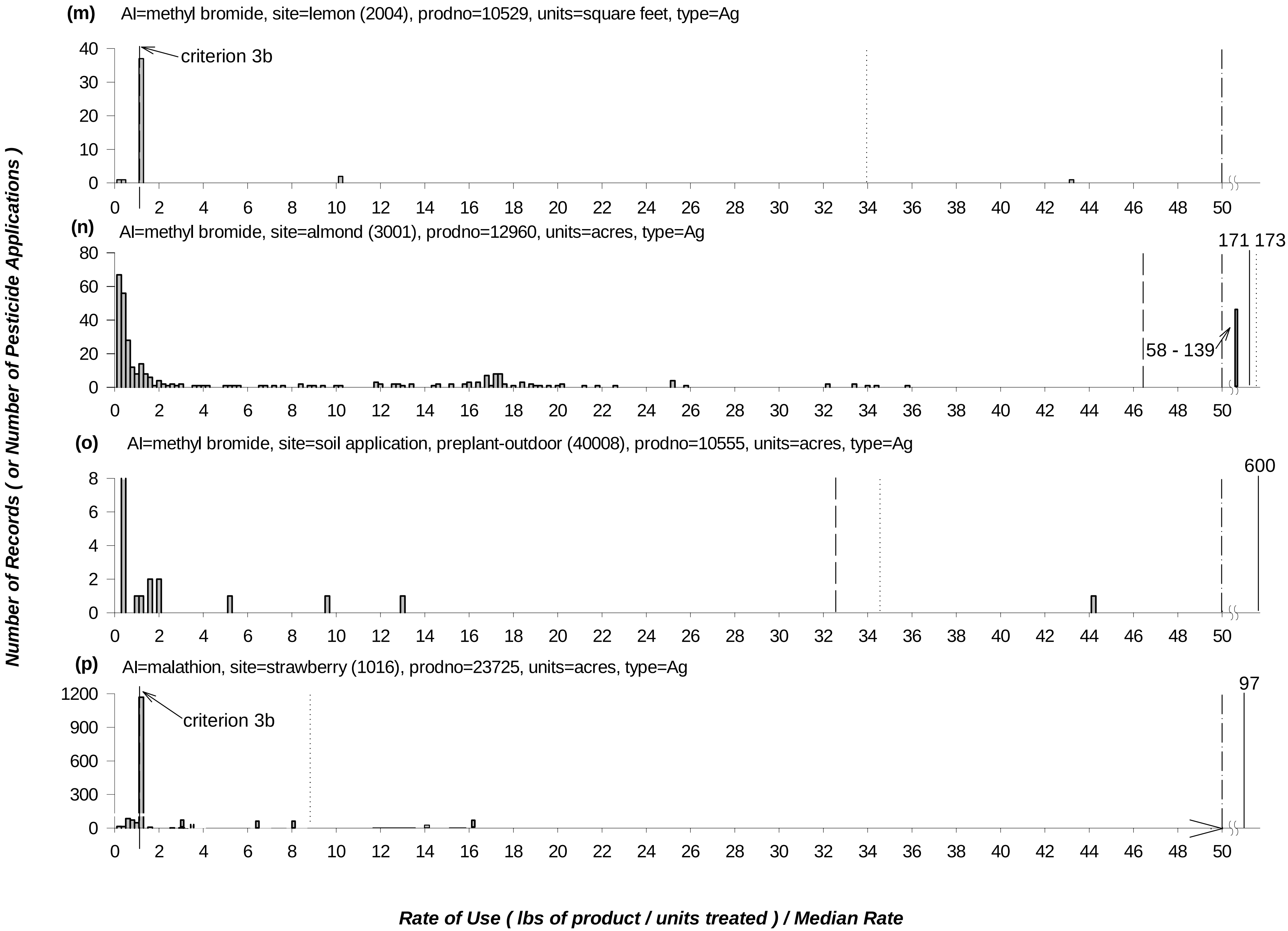
Fig. 1. Frequency distributions of the pesticide rates of use. Each graph gives the number of records (each record is one pesticide application) at which different application rates were used during 1995 in California of a particular use type. A use type is defined by a particular pesticide product, crop/site, unit treated, and record type (that is, agricultural or non-agricultural). Each graph is identified by a letter, the name of the active ingredient in the product, the crop or site (with the PUR site code in parenthesis), the PUR product code, the unit treated, and the record type (agricultural or non-agricultural). To facilitate comparisons between graphs, the use rates were normalized by dividing each rate by the median of all rates for that use type. Four vertical lines were drawn on the plots to mark limit values for the four criteria. The solid line marks the value for criterion 1a, the dashed-dot line for criterion 2b, the dashed line for criterion 3b, and the dotted line for criterion 4d. Some values (of both criteria limits and histograms) were above 50, the maximum normalized rate shown on the graphs.

The actual values for any of the criteria limits above 50 are written above the limit line and the actual values for any histogram are written to the side of the histogram with an arrow pointing to the histogram. If a histogram above 50 represents many rates, either all the rate values are listed or, if there are more than three values, only the range of rate values is given. In graphs with large of number of records, histograms with few records are so small that they are almost impossible to see. These small histograms have been increased in height so that they can be more easily seen, but in most cases they represent only one record.

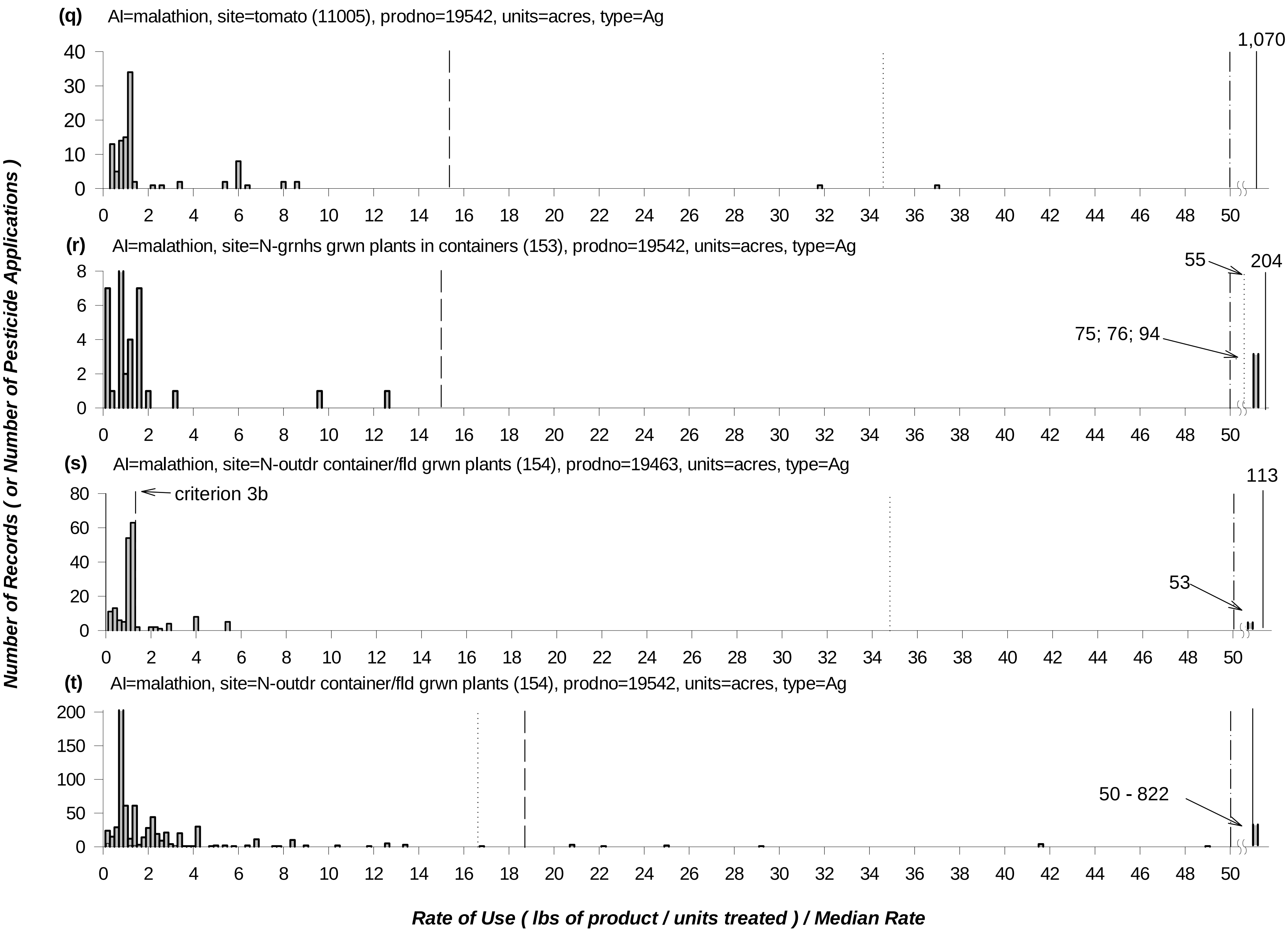

(a)
AI=carbaryl, site=tomato (11005), prodno=1437, units=acres, type=Ag
criterion 1a
criterion 2b
criterion 4d
criterion 3b
134
(b)
AI=carbaryl, site=cabbage (13007), prodno=17559, units=acres, type=Ag
criterion 3b
8,000
60
(c)
AI=carbaryl, site=lemon (2004), prodno=1437, units=acres, type=Ag
1,609
80; 96; 144
(d)
AI=carbaryl, site=orange (2006), prodno=1411, units=acres, type=Ag
Number of Records (or Number of Pesticide Applications)
Rate of Use (lbs of product / units treated) / Median Rate


(e) AI=carbaryl, site=orange (2006), prodno = 1437, units=acres, type=Ag
10,000
(f) AI=carbaryl, site=melons (29122), prodno=8389, units=acres, type=Ag
criterion 3b
200
67
(g) AI=carbaryl, site=tomatoes, for processing/canning (29136), prodno=1412, units=acres, type=Ag
criterion 3b
125
58
(h) AI=carbaryl, site=peppers (8050), prodno=1411, units=acres, type=Ag
criterion 3b
125
100
Number of Records (or Number of Pesticide Applications)
Rate of Use (lbs of product/units treated) / Median Rate

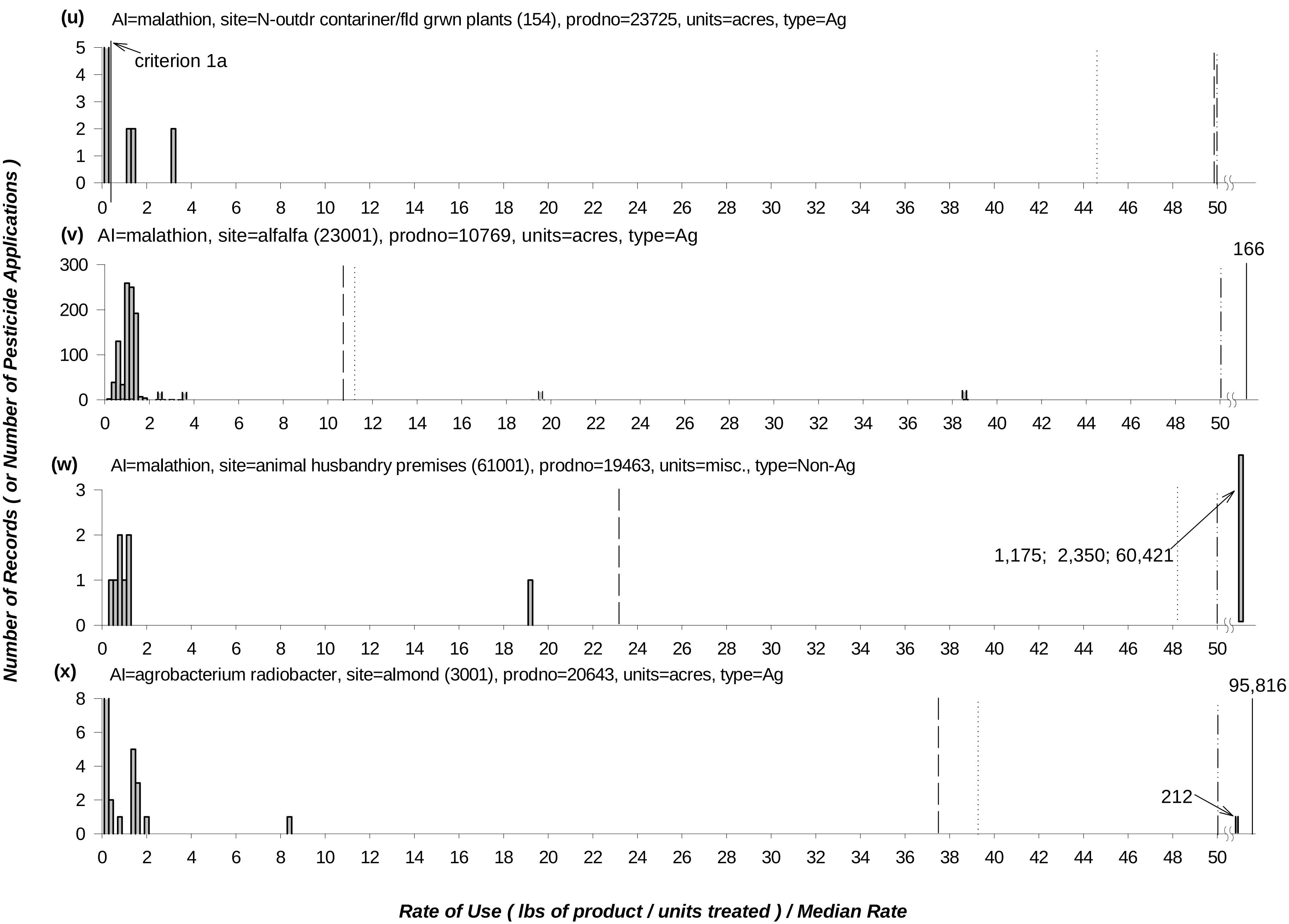

Number of Records (or Number of Pesticide Applications)
(i) AI=metam - sodium, site=carrots (29111), prodno=30047, units=acres, type=Ag
criterion 3b
100
80
60
40
20
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50
(j) AI=methyl bromide, site=strawberry (1016), prodno=10545, units=acres, type=Ag
criterion 3b
50
40
30
20
10
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50
(k) AI=methyl bromide, site=tomato (11005), prodno=16693, units=acres, type=Ag
3
2
1
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50
(l) AI=methyl bromide, site=N-outdr grwn cut flwrs or greens (152), prodno=12960, unit=acres, type=Ag
80
60
40
20
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50
Rate of Use (lbs of product / units treated) / Median Rate



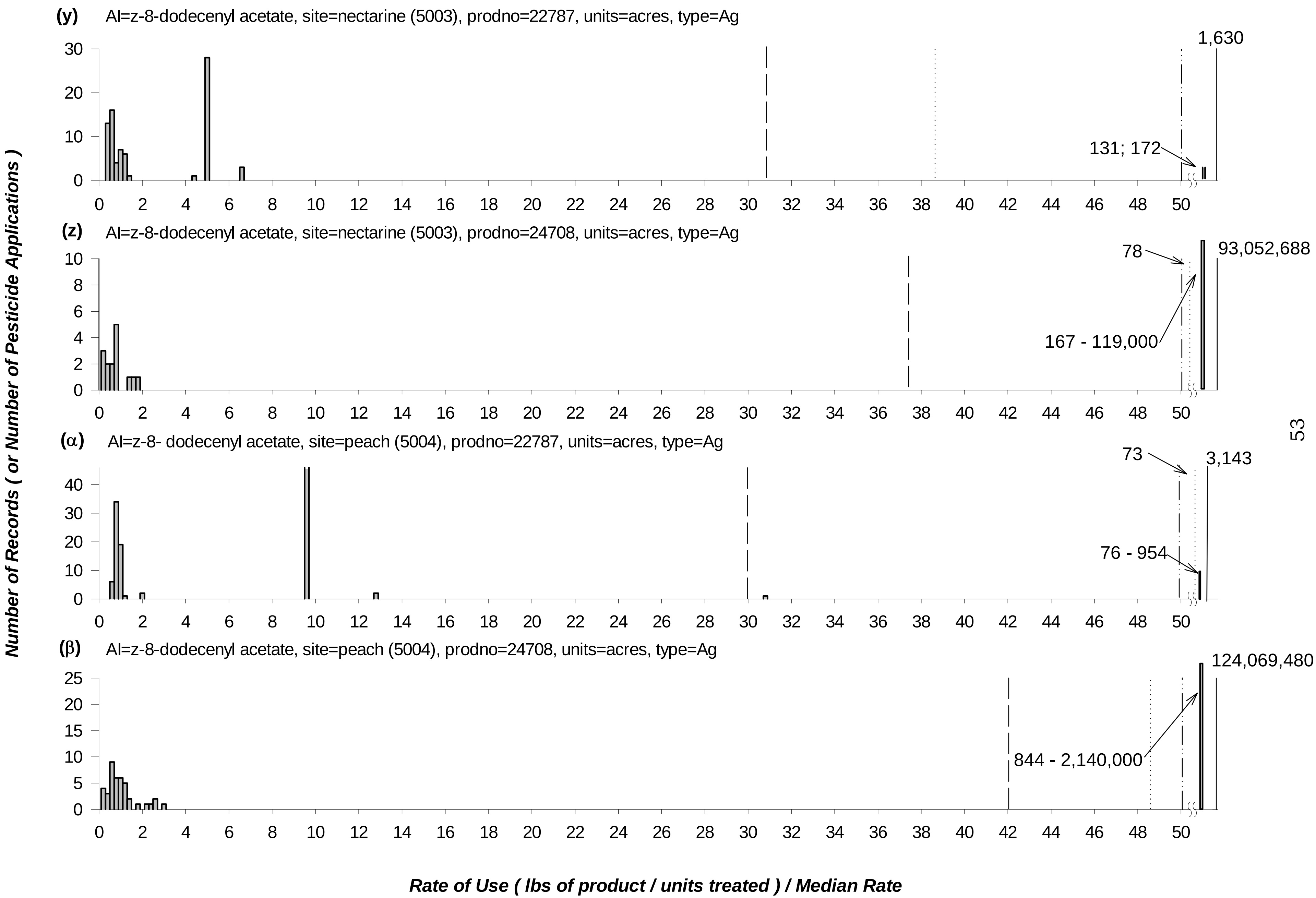
(m) AI=methyl bromide, site=lemon (2004), prodno=10529, units=square feet, type=Ag
criterion 3b
(n) AI=methyl bromide, site=almond (3001), prodno=12960, units=acres, type=Ag
171 173
58 - 139
(o) AI=methyl bromide, site=soil application, preplant-outdoor (40008), prodno=10555, units=acres, type=Ag
600
(p) AI=malathion, site=strawberry (1016), prodno=23725, units=acres, type=Ag
97
criterion 3b
Number of Records (or Number of Pesticide Applications)
Rate of Use (lbs of product / units treated) / Median Rate

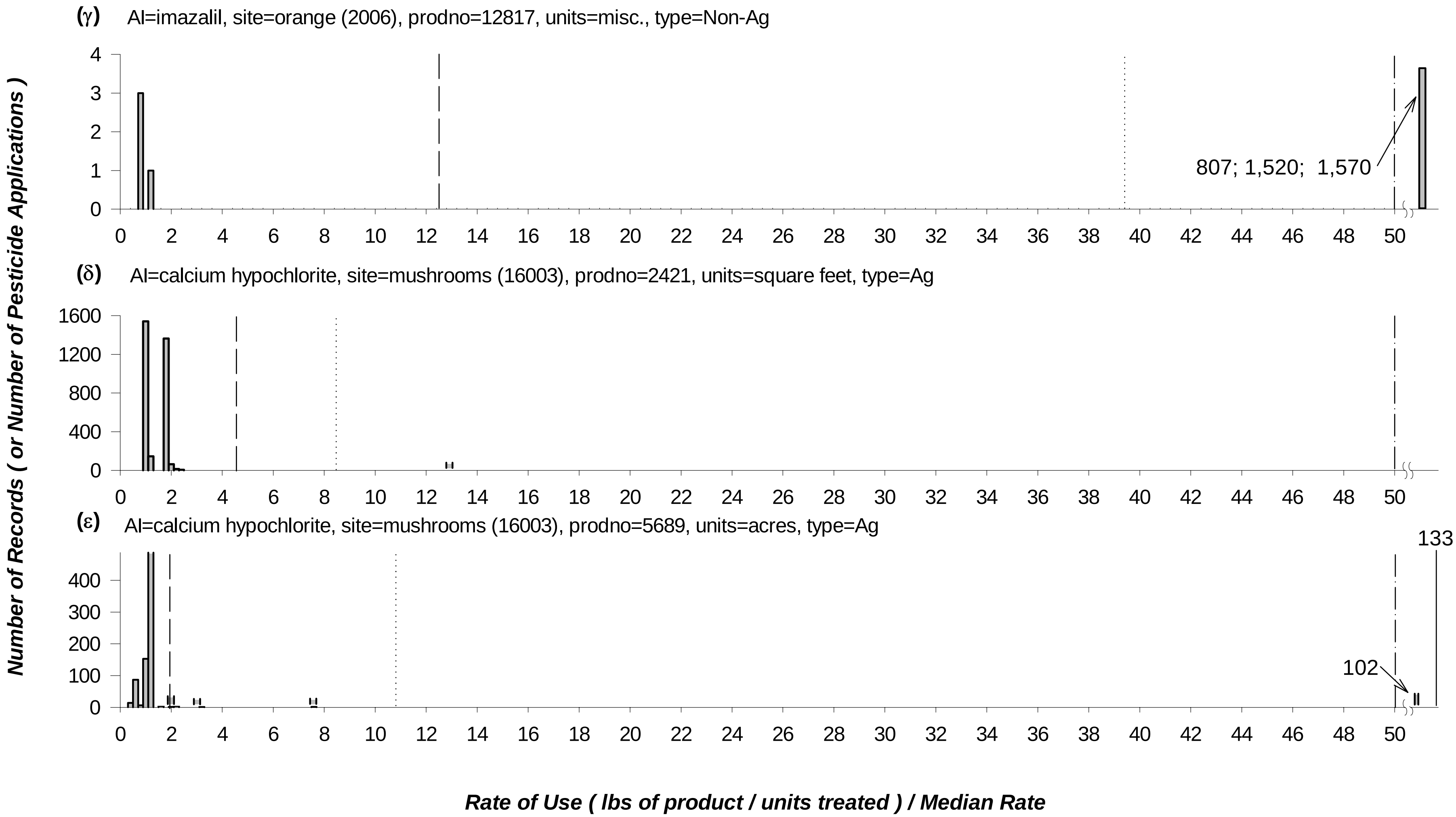


(q) AI=malathion, site=tomato (11005), prodno=19542, units=acres, type=Ag
1,070
(r) AI=malathion, site=N-grnhs grwn plants in containers (153), prodno=19542, units=acres, type=Ag
55
204
75; 76; 94
(s) AI=malathion, site=N-outdr container/fld grwn plants (154), prodno=19463, units=acres, type=Ag
criterion 3b
113
53
(t) AI=malathion, site=N-outdr container/fld grwn plants (154), prodno=19542, units=acres, type=Ag
50 - 822
Number of Records (or Number of Pesticide Applications)
Rate of Use (lbs of product / units treated) / Median Rate



(u) AI=malathion, site=N-outdr contariner/fld grwn plants (154), prodno=23725, units=acres, type=Ag
criterion 1a
(v) AI=malathion, site=alfalfa (23001), prodno=10769, units=acres, type=Ag
166
(w) AI=malathion, site=animal husbandry premises (61001), prodno=19463, units=misc., type=Non-Ag
1,175; 2,350; 60,421
(x) AI=agrobacterium radiobacter, site=almond (3001), prodno=20643, units=acres, type=Ag
95,816
212
Number of Records (or Number of Pesticide Applications)
Rate of Use (lbs of product / units treated) / Median Rate


(y) AI=z-8-dodecenyl acetate, site=nectarine (5003), prodno=22787, units=acres, type=Ag
1,630
131; 172
(z) AI=z-8-dodecenyl acetate, site=nectarine (5003), prodno=24708, units=acres, type=Ag
78
93,052,688
167 - 119,000
(α) AI=z-8- dodecenyl acetate, site=peach (5004), prodno=22787, units=acres, type=Ag
73
3,143
76 - 954
(β) AI=z-8-dodecenyl acetate, site=peach (5004), prodno=24708, units=acres, type=Ag
124,069,480
844 - 2,140,000
Number of Records (or Number of Pesticide Applications)
Rate of Use (lbs of product / units treated) / Median Rate


(γ) AI=imazalil, site=orange (2006), prodno=12817, units=misc., type=Non-Ag
807; 1,520; 1,570
(δ) AI=calcium hypochlorite, site=mushrooms (16003), prodno=2421, units=square feet, type=Ag
(ε) AI=calcium hypochlorite, site=mushrooms (16003), prodno=5689, units=acres, type=Ag
133
102
Number of Records (or Number of Pesticide Applications)
Rate of Use (lbs of product / units treated) / Median Rate

Appendix II: Instructions given to DPR scientists for marking outliers

I am writing a computer program that will flag each record in the Pesticide Use Report in which the reported pounds of pesticide used appears to be an outlying value and thus, presumably, an error. In order to get information needed by the program on what is considered an outlying value, I am asking you, as well as other people, to view the attached set of frequency distributions and mark values you think are outliers.

You should make two judgments of what values are outliers by drawing up to two circles on each graph. One circle should enclose those values in the graphs which you think are almost certainly outliers. The other circle should enclose those values, which you think may be outliers. You should mark only outliers that are extremely large—ignore extremely small values.

The first judgment should be based on the assumption that you need to use the data for some kind of analysis but do not have time to examine the values more closely and must decide whether or not to include some extreme values. Presumably, you would leave out any data which was clearly too extreme to be reasonable and whose inclusion would distort the results. The second judgment should be based on the assumption that you have more time to do an analysis and need results that are as accurate as possible. You might want to more closely examine suspicious values. Thus the second judgment would include in addition to the outliers from the first judgment, other values which appear suspicious. Obviously, you should not choose more suspicious values than you would want to examine. These judgments can be made on the set of plots simply by circling the two groups of outliers. If you make only one circle you should label the circle as identifying obvious outliers or as identifying suspicious outliers—that is, just write “obvious” or “suspicious” by the circle.

Each graph shows the frequency distribution of the pounds of a pesticide product used on a particular site (e.g. crop) per unit area (such as acre or square feet). The horizontal axis represents the pounds of pesticide product per unit area divided by the median value. Thus the median value for all distributions is 1 and a value of 25; for example, means 25 time the median. The vertical axis represents the number of records in which the normalized pounds per unit lies within a 0.2 interval. The graphs all run up to 25 times the median. Any values higher than that are shown in a histogram after the axis break, and the normalized pounds per unit area for all such records are labeled above the histogram. The graphs on the last page have ranges that are larger than 25 since they had a large number of values greater than 25 times the median.

Appendix III: The source code for the outlier program

```
/*
 *         rout.pc
 *         This is an Oracle Pro*C program that generates statistics that are stored in the
 *    Oracle table usetype2002stats.  This table is used by the query rout2002.sql
 *    to flag outliers in the PUR.
 *
 *    To run this program for some year, change all references to the year you want.
 *    For example, if the current file is for year 1998 and you want to run it
 *    for 1999, do a search and replace of '98' to '99'.
 *    However, you need to do this replacement one at a time so that you
 *    do not accidently change a literal numeric value, such as appeaer in the
 *    function Normalize().
 *
 *    You also need to change to the correct password to login into Oracle in main().
 *
 *         This program must be compiled using the make file "proc.mk".
 *    This make file first translates the embedded SQL commands to C funtions
 *    and produces a C source file.
 *    It then uses the Sun Unix C compiler to create the object and executable
 *    files.
 *
 *    To compile this program, type:
 *              make -f proc.mk build EXE=rout2002 OBJS=rout2002.o
 *              (or use the script named "mkmk")
 *
 *
 *******************************************************************************************
 *    The program reads an ORACLE PURXX table and adds a 'Y' or 'N' to several
 *         different table fields in the table OUTLIERXX (where XX is the year)
 *         to flag outlier values based on different sets of criteria.
 *         If a value is an outlier, a 'Y' is placed in the field, if not,
 *         'N' if placed in the field.  If the value in the field acre_treated is 0,
 *         or if a decision cannot be made for a record, the outlier fields are left
 *         blank.  Statistics and outlier results are stored in a table named
 *         OUTLIERXX_STATS.  Both tables, OUTLIERXX and OUTLIERXX_STATS,
 *         must be already created with the correct fields.
 *         The OUTLIERXX table must contain the following VARCHAR2(1) fields
 *         (with a description of the outlier criteria used):
 *
 *         ai_a_1000_200--flagged if the pounds per acre treated of any of the AIs
 *              in the product is greater than MAX_LBS_AI (set now to 200), unless
 *              it is a high use pesticide (methyl bromide, chloropicrin, metam-sodium,
 *              dazomet, boric acid, or 1,3 dichloropropene) which must be greater than
 *              MAX_LBS_AI_HUP (1000).
 *              Note: only records are flagged in which units treated are acres.
 *
 *         ai_a_2000_400--same as previous field except MAX_LBS_AI and MAX_LBS_AI_HUP
 *              have been doubled.
 *
 *    prd_u_25m--flagged if the pounds per units treated of the product
 *              is greater than RANGE1 (25) times the median value for all uses of
 *              that product on the same site, the same unit treated, and the same
 *              record type (see below for meaning of record type). Units treated can
 *              be acres, square feet, cubic feet, or other measures.
 *
 *    prd_u_50m--same as previous field except use RANGE2 (50) rather than RANGE1.
 *
 *    prd_u_10md--flagged if the pounds per units treated of the product is
 *              greater than the median + MEDDEV1 (10) times the median difference.
 *              Flagged only if the there are >= 50 records with same product, site_code,
 *              unit treated, and record type and only if a small proportion of these
 *              are possible outliers. (A small proportion is < 5% of the number of records
 *              in this group when number of records is <= 200, or < 2% if the number of
 *              records is > 200 and <= 1000, or < 1% if the number of records > 1000).
 *              Median is the same as in previous field and median difference is the
 *              median of all the absolute values of the differences between the
 *              median and each individual value.
 *
 *    prd_u_50md--same as previous field except use MEDDEV2 (50) rather than MEDDEV1.
 *
```

```
 *          acre700--flagged if the number of acres treated is greater than 700.
 *
 *          There are two different kinds of record_id's (one of the fields in the PUR).
 *          Record type refers to one of the kinds of record_id's.  Record_id's
 *          of C, 2, G, D are for structural, rights of way, etc.  They are not location
 *          specific and each record can be the sum of many applications.  The
 *          other record_id's (A, B, 1, 4, E, F) are agricultural sites.
 *
 *          To run the program for a different PUR table, the name of the table must
 *          changed throughout this source file and then recompiled.
 */

#include <stdio.h>
#include <math.h>
#include <string.h>
#include <sqlca.h>

#define MAX_LBS_AI            200.0    /* Maximum allowable lbs of AI per acre */
#define MAX_LBS_AI_HUP  1000.0   /* Maximum allowable lbs of high use AI/acre */
#define RANGE1                25.0        /* First maximum value times the median */
#define RANGE2                50.0     /* Second maximum value times the median */
#define MEDDEV1               10.0        /* First maximum deviation from the median */
#define MEDDEV2               50.0     /* Second maximum deviation from the median */
#define NUM_RECORDS     32767    /* Note: array size must be < 32K (Oracle limitation) */
#define NUM_CHEMS             20       /* Max number of AIs in a pesticide product */
#define NUM_STATS             68       /* Number of statistical measures for neural network
                                                                       plus one extra for
threshold value */
#define NUM_HIDDEN            67       /* Number of nodes in hidden layer of neural network */
#define NUM_NNVALUES          4        /* Number of neural network outlier values (output from
NN) */
#define NUM_OUTVALUES   6        /* Number of other outlier values */

#include "weights.c"             /* Weights for the neural network. In an included source
                                                                       file just to reduce the
distracting mass of numbers. */

void CalculateStats();
void GetStats2(int numRecs, double *lbs_array, long prodno, char *unit_treated,
      double *median, double *med_dev,
      double *outvalues, double *nnvalues, short *out_ind, short *nn_ind);
void NeuralNet(const double *inputs, double *outputs);
double GetOutlierLbsProduct(long prodno);
void Stats(int n, const double data[], double stats[]);
void StatsTrim(int n, const double data[], double stats[]);
double Sigmoid(const double x);
void Normalize(const double data[], double norm[]);
void UnNormalize(double data[], const double norm[], const double median);
double Median(int numRecs, const double *values);
double MedDev(int numRecs, const double *values, double median);
void Sort(int n, double *ra);
void sql_error(char *msg);
void Error(char *str1, char *str2, char *str3);

/* FILE *fp; */

main()
{
   /*
   char *connect = "/";
   */

      char *username = "pur";
      char *password = "";

      /* Connect to ORACLE. */
       EXEC SQL WHENEVER SQLERROR DO sql_error("Connect error:");

       /* EXEC SQL CONNECT :connect;
     */
```

```
    EXEC SQL CONNECT :username IDENTIFIED BY :password;
       printf("\nConnected to ORACLE as user: %s\n", username);

       EXEC SQL WHENEVER SQLERROR DO sql_error("Oracle error:");
       EXEC SQL SET TRANSACTION USE ROLLBACK SEGMENT R_LARGE;

   /* Calculate all statistics and store in table outlierXX_stats
    */
    CalculateStats();

       printf("\nAu revoir-rout2.\n\n\n");
    /* fclose(fp); */

      /* Disconnect from the database. */
       EXEC SQL COMMIT WORK RELEASE;
       exit(0);
}

/*
 * Find statistics for current record.
 */
void CalculateStats()
{
      long prodno;
   long site_code;
   char unit_treated[2];
   char record_id_type[2];

   char buf[80];
      double lbs_array[NUM_RECORDS];
   int numRecs;
   int i;
   double median, med_dev;
   double nnvalues[NUM_NNVALUES];
   double outvalues[NUM_OUTVALUES];
   short out_ind[NUM_OUTVALUES];
   short nn_ind[NUM_NNVALUES];
   short prodno_ind, site_code_ind, unit_treated_ind, record_id_type_ind;

   /* Create cursor to retrieve each row of the table usetypeXX --
    * The usetypeXX table was created previously to hold all distinct
    * use type values.  This table will here be filled with the
    * criteria limit values.
    */
   EXEC SQL
   DECLARE usetype_cursor CURSOR FOR
      SELECT prodno, site_code, unit_treated, record_id_type
      FROM usetype2002;

   EXEC SQL OPEN usetype_cursor;

      for( ; ; ) {
            EXEC SQL WHENEVER NOT FOUND DO break;

            /* Get the next use type */
            EXEC SQL FETCH usetype_cursor
                  INTO  :prodno:prodno_ind, :site_code:site_code_ind,
                  :unit_treated:unit_treated_ind, :record_id_type:record_id_type_ind;

            EXEC SQL WHENEVER NOT FOUND CONTINUE;

      EXEC SQL
      SELECT      NVL(LBS_PRD_USED/ACRE_TREATED, 0)
      INTO        :lbs_array
      FROM        purrates2002
      WHERE       acre_treated > 0  AND
               (prodno = :prodno OR
                  ((prodno IS NULL) AND (:prodno:prodno_ind IS NULL))) AND
               (site_code = :site_code OR
                  ((site_code IS NULL) AND (:site_code:site_code_ind IS NULL))) AND
```

```
               (unit_treated = :unit_treated OR
                  ((unit_treated IS NULL) AND (:unit_treated:unit_treated_ind IS NULL))) AND
               (record_id_type = :record_id_type OR
                  ((record_id_type IS NULL) AND (:record_id_type:record_id_type_ind IS NULL)));

      numRecs = sqlca.sqlerrd[2];
      if( numRecs > NUM_RECORDS )  {
         sprintf(buf, "Too many records for site_code %ld, prodno = %ld", site_code, prodno);
         Error( buf, "", "");
      }
      Sort( numRecs, lbs_array );
      GetStats2(numRecs, lbs_array, prodno, unit_treated, &median, &med_dev, outvalues,
nnvalues, out_ind, nn_ind);

            /* Store statistical values in the outlier table */
            EXEC SQL
      INSERT INTO usetype2002stats
      VALUES(  :prodno:prodno_ind, :site_code:site_code_ind, :unit_treated:unit_treated_ind,
                       :record_id_type:record_id_type_ind, :numRecs, :median, :med_dev,
                       :outvalues[0]:out_ind[0], :outvalues[1]:out_ind[1],
                 :outvalues[2]:out_ind[2], :outvalues[3]:out_ind[3],
                 :outvalues[4]:out_ind[4], :outvalues[5]:out_ind[5],
                 :nnvalues[0]:nn_ind[0], :nnvalues[1]:nn_ind[1],
                 :nnvalues[2]:nn_ind[2], :nnvalues[3]:nn_ind[3]);
      EXEC SQL COMMIT;
      EXEC SQL SET TRANSACTION USE ROLLBACK SEGMENT R_LARGE;
   }

   EXEC SQL CLOSE usetype_cursor;
}


/*
 * Find statistics for current record.
 */
void GetStats2(int numRecs, double *lbs_array, long prodno, char *unit_treated,
      double *median, double *med_dev,
      double *outvalues, double *nnvalues, short *out_ind, short *nn_ind)
{
      double stats[NUM_STATS];

   if( numRecs > 1 ) {
      Stats(numRecs, lbs_array, stats);
      NeuralNet(stats, nnvalues);
      *median = stats[1];
      *med_dev = stats[2];
   } else {
      stats[0] = numRecs;
      stats[1] = lbs_array[0];
   }

   /* Criteria 1 and 2 (outvalues[0] and outvalues[1]) are only used when units are acres.
      If units are not in acres, the associated indicator variables equal -1 */
   if( unit_treated[0] == 'A' )  {
      outvalues[0] = GetOutlierLbsProduct(prodno);
      outvalues[1] = 2.0*outvalues[0];
      out_ind[0] = out_ind[1] = 0;
   } else {
      out_ind[0] = out_ind[1] = -1;
   }

   /* Criteria 3 and 4 require more than one record */
   if( numRecs > 1) {
      outvalues[2] = RANGE1*(*median);
      outvalues[3] = RANGE2*(*median);
      out_ind[2] = out_ind[3] = 0;
   } else {
      out_ind[2] = out_ind[3] = -1;
   }
```

```
   /* Criteria 5 and 6 require more than two records
      Note: in previous version 0.001 was added to these limit values */
   if( numRecs > 2) {
      outvalues[4] = (*median) + MEDDEV1*(*med_dev);
      outvalues[5] = (*median) + MEDDEV2*(*med_dev);
      out_ind[4] = out_ind[5] = 0;
   } else {
      out_ind[4] = out_ind[5] = -1;
   }

   /* Criteria 7 to 10 require more than one record and positive outlier limits
    * Ideally, here I would place nulls into nnvalues if any of the nnvalues
    * were 0, but to do this I would then need to check for null values in the
    * criteria 7 to 10 statements in FlagOutliers().
    */
   if( numRecs > 1 ) {
      nn_ind[0] = nn_ind[1] = nn_ind[2] = nn_ind[3] = 0;
   } else {
      nn_ind[0] = nn_ind[1] = nn_ind[2] = nn_ind[3] = -1;
   }
}

void NeuralNet(const double *inputs, double *outputs)
{

      double ninputs[NUM_STATS];    /* Normalized inputs */
   double hidden[NUM_HIDDEN+1];
   double noutputs[NUM_NNVALUES];  /* Normalized outputs */
   double wx, wh;
   int i, j;

   Normalize(inputs, ninputs);

   for(i=0; i<NUM_HIDDEN; i++) {
      wx = 0.0;
      for(j=0; j<NUM_STATS; j++)
         wx += weightsIn[i][j]*ninputs[j];
      hidden[i] = Sigmoid(wx);
   }
   hidden[NUM_HIDDEN] = 1.0;

   for(i=0; i<NUM_NNVALUES; i++) {
      wh = 0.0;
      for(j=0; j<NUM_HIDDEN+1; j++)
         wh += weightsOut[i][j]*hidden[j];
      noutputs[i] = Sigmoid(wh);
   }

   UnNormalize(outputs, noutputs, inputs[1]);
}

/*
 * Find the pounds of product per acre treated in which the either the
 * pounds/acre of non-high use AI's in the product equals 200 or pounds/acre
 * of the high use AI's in the product equals 1000.
 */
double GetOutlierLbsProduct(long prodno)
{
      int chem_codes[NUM_CHEMS];
      double prodchem_pcts[NUM_CHEMS];
      double max_prodchem_pct, max_prodchem_pct_hup;
   double maxLbsProd, maxLbsProdHup;
      int numRecs;     /* number of rows returned */
      int j;

      /* Get the chemical codes of each AI in the product and the percentages of each
            in the product */
      EXEC SQL SELECT CHEM_CODE, PRODCHEM_PCT
            INTO :chem_codes, :prodchem_pcts
            FROM  PROD_CHEM
```

```
            WHERE PRODNO = :prodno AND CHEM_CODE > 0;

      numRecs = sqlca.sqlerrd[2] < NUM_CHEMS ? sqlca.sqlerrd[2] : NUM_CHEMS;

   if(numRecs == 0)
      return 0.0;

      /* Find the AI in the product with the highest percentage.
            If one or more of the AIs is a "high use pesticide" then find which one
            has the highest percentage and, in addition find which of the other
            AIs has the highest percentage */
      max_prodchem_pct = 0.0;     /* The highest percent AI (not high use AI) in product */
      max_prodchem_pct_hup = 0.0; /* The highest percent AI (high use AI) in product */
   /* fprintf(fp, "prod = %ld; prodpcts: ", prodno); */
      for(j=0; j<numRecs; j++) {
            if( chem_codes[j] == 385 || chem_codes[j] == 136 || chem_codes[j] == 573 ||
             chem_codes[j] == 233 || chem_codes[j] == 769 || chem_codes[j] == 616 )  {
                  max_prodchem_pct_hup = max_prodchem_pct_hup > prodchem_pcts[j] ?
                        max_prodchem_pct_hup : prodchem_pcts[j];
            } else {
                  max_prodchem_pct = max_prodchem_pct > prodchem_pcts[j] ?
                        max_prodchem_pct : prodchem_pcts[j];
            }
      /* fprintf(fp, "%5.2f ", prodchem_pcts[j]); */
      }

      maxLbsProd = max_prodchem_pct > 0.0 ? MAX_LBS_AI*100/max_prodchem_pct : -1.0;
      maxLbsProdHup = max_prodchem_pct_hup > 0.0 ? MAX_LBS_AI_HUP*100/max_prodchem_pct_hup : -
1.0;

   /* fprintf(fp, "Max lbs: %6.2f %6.2f \n", maxLbsProd, maxLbsProdHup); */

   /* Return the minimum of the two outlier values for pounds of product.
      Thus, if the pounds of product is greater than either one of these
      values, then the record is flagged as an outlier. */

   if(  maxLbsProd < 0.0  && maxLbsProdHup < 0.0 )
      return 0.0;

   if( maxLbsProd < 0.0 )
      return maxLbsProdHup;

   if( maxLbsProdHup < 0.0 )
      return maxLbsProd;

   return maxLbsProd < maxLbsProdHup ?  maxLbsProd : maxLbsProdHup;
}

void Stats(int n, const double data[], double stats[])
{
   double median;
      int i;
   double sum = 0.0;
   double p, d;
   double mean, var = 0.0, stddev, skew=0.0, kurt=0.0;

   if(n<1)
      Error("Must have at least 1 value in Stats()", "", "");

      median = Median(n, data);
      /* Statistics used by the neural net are calculated on median normalized
      data.  However, to make it easier to use the statistics for other purposes
      statistics are calculated normally here. The statistical values
      are later transformed in the neural network function */

   for( i=0; i<n; i++ )
      sum += data[i];

   mean = sum/n;
```

```
   for( i=0; i<n; i++ ) {
      d = data[i] - mean;
      var += (p = d*d);
      skew += (p *= d);
      kurt += (p *= d);
   }

   if( n > 0 ) {
      var /= (n-1);
      stddev = sqrt(var);
   } else {
      var = 0.0;
      stddev = 0.0;
   }

   if( var && n > 3) {
      skew *= n/((n-1)*(n-2)*var*stddev);
      kurt = kurt*n*(n+1)/((n-1)*(n-2)*(n-3)*var*var) - 3.0*(n-1)*(n-1)/((n-2)*(n-3));
   } else {
      skew = kurt = 0.0;
   }

   stats[0] = (double)n;
   stats[1] = median;
   stats[2] = MedDev(n, data, median);
   stats[3] = mean;
   stats[4] = stddev;
   stats[5] = skew;
   stats[6] = kurt;
   StatsTrim(n, data, &stats[7]);
   stats[NUM_STATS-1] = 1.0;
}

void StatsTrim(int n, const double data[], double stats[])
{
      int reduce, numRows, i, j;
   double sum, sumsq, sum3, sum4, dat, d, var, stddev;

   for(i=0; i<60; i++)
      stats[i] = 0.0;

   if(n<2)   return;

   reduce = n/100 + 1;
   numRows = n - reduce*16;

   sum = sumsq = sum3 = sum4 = 0.0;
   for(i=0; i<numRows; i++) {
      dat = data[i];
      sum += (d = dat);
      sumsq += (d *= dat);
      sum3 += (d *= dat);
      sum4 += (d *= dat);
   }

   for(i=15; i>=1; i--) {
      numRows += reduce;
      if(numRows > 0 ) {
         for(j=0; j<reduce; j++) {
            dat = data[numRows+j-reduce];
            sum += (d = dat);
            sumsq += (d *= dat);
            sum3 += (d *= dat);
            sum4 += (d *= dat);
         }
            stats[i-1] = sum/numRows;                                    /* mean */
         if(numRows > 1) {
                  var = (sumsq - sum*sum/numRows)/(numRows-1);    /* var */
                  stats[15+i-1] = stddev = sqrt( var );                 /* standard deviation */
            if(numRows > 2 && var > 0 ) {                     /*skewness */
```

```
                        stats[30+i-1] = (numRows*sum3 - 3.0*sum*sumsq + 2.0
*sum*sum*sum/numRows)/
                         ( (numRows - 1)*(numRows - 2)* var*stddev);
               if(numRows > 3)                              /* kurtosis */
                              stats[45+i-1] = (numRows+1)*(numRows*sum4 - 4.0*sum*sum3 + 6.0
*sum*sum*sumsq/numRows - 3.0*sum*sum*sum*sum/(numRows*numRows))/
                              ( (numRows - 1)*(numRows - 2)*(numRows - 3)* var*var) -
                                                3.0*(numRows - 1)*(numRows - 1)/((numRows - 2)*
(numRows - 3));
            }
         }
      }
   }
}

double Sigmoid(const double x)
{
      if(x > 20.0) return 1.0;
   if(x < -20.0) return 0.0;
   return 1.0/(1.0 + exp(-x));
}

void Normalize(const double data[], double norm[])
{
   static double min[NUM_STATS] =
      { 0.00001,5,0.59617,0,0,-15.17,-
2.2425,0.58271,0.53054,0.08456,0.06108,0,0,0,0,0,0,0,0,0,0,0,0,0,
      0,0,0,0,0,0,0,0,0,0,0,0,-15.174,-15.099,-15.02,-14.939,-14.858,-14.776,-14.694,-14.611,-
14.528,
      -14.444,-14.36,-14.276,-14.19,-14.105,-14.019,-2.2499,-2.4824,-2.9444,-3.3333,-6,-6,-
3.3326,
      -5.9969,-5.9991,-2.3076,-3.308,-5.9659,-3.0796,-3.8766,-3.3333,0.0 };
   static double max[NUM_STATS] =
      {
1000,1823,5698.27,25653.6,0.93891,33.4791,1264.96,4178.56,2720.91,1225.68,15.5645,13.7258,
      11.7904,10.4137,9.2485,8.01664,6.71231,5.46581,4.19034,3.6217,3.05367,2.74794,21903.9,

17785.9,12010.2,25.1378,22.5776,19.3258,17.6024,16.3407,14.7864,12.8029,10.7032,7.92698,7.36474,

6.65669,5.7299,14.0007,8.29336,9.73356,7.66644,9.53491,7.14114,4.31597,5.04643,6.25328,8.92048,

12.7536,4.12433,5.55035,4.17719,4.45958,258.124,255.535,252.843,250.128,247.41,244.692,241.974,
      239.256,236.538,233.82,231.102,228.384,225.666,222.949,220.231,1.0 };
   double medianNormData;
      int j;

   /* For the neural network, the data used were normalized by dividing by
      the median.  Values of skewness and kurtosis are not affected by this
      transformation, but mean, and standard deviations are, so these statistical
      values need to divided by the median as well.

      All values need to be further normalized by using max and min values. */

   for(j=0; j<NUM_STATS; j++) {
      if( (j >=2 && j <= 4) || (j >=7 && j <= 36) )
            medianNormData = data[j]/data[1];
      else
            medianNormData = data[j];

      if( medianNormData <= min[j] )
         norm[j] =  0.0;
      else if( medianNormData >= max[j] )
         norm[j] = 1.0;
      else
         norm[j] =  (medianNormData - min[j])/(max[j] - min[j]);
   }
}

void UnNormalize(double data[], const double norm[], const double median)
{
```

```
      static double min[NUM_NNVALUES] = { 1.4, 2.5, 5, 8 };
   static double max[NUM_NNVALUES] = { 55, 50, 130, 200 };

   int i;

   for(i=0; i<NUM_NNVALUES; i++)
      data[i] = median*(norm[i]*(max[i] - min[i]) + min[i]);
}

/*
 * Calculate the median of a set of numbers stored in the sorted array "values[]"
 */
double Median(int numRecs, const double *values)
{
      int n2, n2m;

      if( numRecs <= 0 )
            return 0.0;
      if( numRecs == 1)
            return values[0];
      if( numRecs == 2 )
            return 0.5*(values[0] + values[1]);

      n2m = (n2=numRecs/2) - 1;
      return (numRecs % 2 ? values[n2] : 0.5*(values[n2m] + values[n2]) );
}

/*
 * Calculate the median difference of a set of numbers stored in the array values[],
 * whose median value is in the variable "median".
 */
double MedDev(int numRecs, const double *values, double median)
{
      double *diffs;
      double meddiff;
      int i;

      if( numRecs <= 0 )
            return 0.0;

      diffs = (double *)malloc((unsigned)numRecs*sizeof(double));

      if(!diffs)
            Error("Out of memory error in MedDiff()", "", "");

      for(i=0; i<numRecs; i++)
            diffs[i] = fabs(values[i] - median);

      Sort(numRecs, diffs);
      meddiff = Median(numRecs, diffs);
      free((char*)diffs);
      return meddiff;
}

/*
 * Sort the numbers in the array "ra[]".  The number of elements in the
 * is "n".
 * This code is from the book "Numerical Recipes in C" by Press et al. 1988
 */
void Sort(int n, double *ra)
{
      int l, j, ir, i;
      double rra;

      if(n==0 || n==1)
            return;

      l = (n >> 1) + 1;
      ir = n;
      for(;;) {
```

```
            if( l > 1 )
                  rra = ra[--l-1];
            else {
                  rra = ra[ir-1];
                  ra[ir-1] = ra[1-1];
                  if( --ir == 1) {
                        ra[1-1] = rra;
                        return;
                  }
            }
            i = l;
            j = l << 1;
            while( j <= ir ) {
                  if( j < ir && ra[j-1] < ra[j]) j++;
                  if( rra < ra[j-1]) {
                        ra[i-1] = ra[j-1];
                        j += (i=j);
                  }
                  else
                        j = ir + 1;
            }
            ra[i-1] = rra;
      }
}

/*
 * Handles Oracle unrecoverable errors
 */
void sql_error(char *msg)
{
      char buffer[510];
      int bufSize = 510;
      int msgLen;

      EXEC SQL WHENEVER SQLERROR CONTINUE;

      sqlglm(buffer, &bufSize, &msgLen);
      buffer[msgLen] = '\0';
      printf("\n%s", msg);
      printf("\n%s\n", buffer);

      EXEC SQL ROLLBACK WORK RELEASE;
      exit(1);
}

/*
 * Handles all other errors
 */
void Error(char *str1, char *str2, char *str3)
{
      printf("Run-time error...\n");
      printf("%s %s %s\n", str1, str2, str3);
      EXEC SQL ROLLBACK WORK RELEASE;
      exit(1);
}
```

```
/*
 * This query creates tables for flagging outliers in rates of use in the PUR.
 *
 * To run this program for some year, change all references to the year you want.
 * For example, if the current file is for year 2002 and you want to run it
 * for 2003, do a search and replace of '2002' to '2003'.
 *
 * This query calls a C program that generates statistics needed to identify
 * outliers.  These statistics are stored in the table usetype2002stats.
 *
 * The C program needs to be compiled before this query is run.
 * The program is actually an Oracle Pro*C program, in the file "rout2002.pc"
 * See the notes in that file on how to compile the program.
 *
 */
set termout on
set serveroutput on
set document off
SET TRANSACTION USE ROLLBACK SEGMENT R_LARGE;

CREATE TABLE usetype2002
      pctfree 5
      pctused 90
      storage (initial 10M next 2M)
   tablespace PUR
      AS SELECT   DISTINCT prodno, site_code, unit_treated,
                        TRANSLATE(record_id, 'C2G9DHAB14EF', 'NNNNNNAAAAAA') record_id_type
      FROM        pur2002
      WHERE       acre_treated > 0;

/* Table of all use types, some statistics, and the outlier limits for each
 * use type.  These values are calculated in the C program rout2
 */

CREATE TABLE usetype2002stats
   (prodno  NUMBER(7),
    site_code  NUMBER(6),
    unit_treated VARCHAR(1),
    record_id_type VARCHAR(1),
    numrecs NUMBER(10),
    median FLOAT(30),
    med_dev FLOAT(30),
    ai_a_1000_200 FLOAT(30),
    ai_a_2000_400 FLOAT(30),
    prd_u_25M FLOAT(30),
    prd_u_50M FLOAT(30),
    prd_u_10MD FLOAT(30),
    prd_u_50MD FLOAT(30),
    nn1 FLOAT(30),
    nn2 FLOAT(30),
    nn3 FLOAT(30),
    nn4 FLOAT(30))
   pctfree 5
      pctused 90
      storage (initial 10M next 2M)
   tablespace PUR;

GRANT SELECT ON usetype2002stats TO PUBLIC;
CREATE PUBLIC SYNONYM usetype2002stats FOR usetype2002stats;


/* Table with subset of pur2002 fields.
 */

CREATE TABLE purrates2002
      pctfree 5
      pctused 90
      storage (initial 10M next 5M)
   tablespace PUR
      AS SELECT   use_no, prodno, site_code, unit_treated,
```

```
                        TRANSLATE(record_id, 'C2G9DHAB14EF', 'NNNNNNNAAAAAA') record_id_type,
               lbs_prd_used, acre_treated
      FROM        pur2002;

CREATE INDEX purrates2002_ndx ON purrates2002
            (acre_treated, prodno, site_code, unit_treated, record_id_type)
      pctfree 5
      storage( initial 5M next 5M pctincrease 0)
      TABLESPACE NDX;

/*
CREATE INDEX pur2002_psur_ndx ON pur2002
            (prodno, site_code, unit_treated, record_id)
      pctfree 5
      storage( initial 5M next 5M pctincrease 0)
      TABLESPACE NDX;
*/


/* The C program rout2 fills in the usetype2002stats table
 */
host rout2002 > rout2002.cout;

COMMIT;

CREATE INDEX usetype2002stats_ndx ON usetype2002stats
            (prodno, site_code, unit_treated, record_id_type)
      pctfree 5
      storage( initial 5M next 5M pctincrease 0)
      TABLESPACE NDX;

CREATE TABLE OUTLIER2002
            (USE_NO NUMBER(8),
       AI_A_1000_200 VARCHAR2(1),
       AI_A_2000_400 VARCHAR2(1),
       PRD_U_25M VARCHAR2(1),
       PRD_U_50M VARCHAR2(1),
       PRD_U_10MD VARCHAR2(1),
       PRD_U_50MD VARCHAR2(1),
       NN1 VARCHAR2(1),
       NN2 VARCHAR2(1),
       NN3 VARCHAR2(1),
       NN4 VARCHAR2(1),
       ACRE700 VARCHAR2(1)
    )
pctfree 5
pctused 90
storage( initial 20M next 5M pctincrease 0)
tablespace PUR;

set termout on
set serveroutput on

/* Flag each record in PUR if it is outlier by each criteria
 */
DECLARE
   lbs_per_unit FLOAT(30);

   numrecs usetype2002stats.numrecs%TYPE;
   ai_a_1000_200 usetype2002stats.ai_a_1000_200%TYPE;
   ai_a_2000_400 usetype2002stats.ai_a_2000_400%TYPE;
   prd_u_25m usetype2002stats.prd_u_25m%TYPE;
   prd_u_50m usetype2002stats.prd_u_50m%TYPE;
   prd_u_10md usetype2002stats.prd_u_10md%TYPE;
   prd_u_50md usetype2002stats.prd_u_50md%TYPE;
   nn1 usetype2002stats.nn1%TYPE;
   nn2 usetype2002stats.nn2%TYPE;
   nn3 usetype2002stats.nn3%TYPE;
   nn4 usetype2002stats.nn4%TYPE;
```

```
   ai_a_1000_200_flag VARCHAR2(1);
   ai_a_2000_400_flag VARCHAR2(1);
   prd_u_25m_flag VARCHAR2(1);
   prd_u_50m_flag VARCHAR2(1);
   prd_u_10md_flag VARCHAR2(1);
   prd_u_50md_flag VARCHAR2(1);
   nn1_flag VARCHAR2(1);
   nn2_flag VARCHAR2(1);
   nn3_flag VARCHAR2(1);
   nn4_flag VARCHAR2(1);
   acre700_flag VARCHAR2(1);

   CURSOR pur_cursor  IS
      SELECT      use_no, prodno, site_code, unit_treated,
            TRANSLATE(record_id, 'C2G9DHAB14EF', 'NNNNNNAAAAAA') record_id_type,
            lbs_prd_used, acre_treated
      FROM        pur2002
      WHERE       acre_treated > 0;
BEGIN
   FOR pur_rec IN pur_cursor LOOP
      BEGIN
         SELECT   numrecs, ai_a_1000_200, ai_a_2000_400, prd_u_25m, prd_u_50m,
                         prd_u_10md, prd_u_50md, nn1, nn2, nn3, nn4
         INTO           numrecs, ai_a_1000_200, ai_a_2000_400, prd_u_25m, prd_u_50m, prd_u_10md,
                        prd_u_50md, nn1, nn2, nn3, nn4
         FROM           usetype2002stats
         WHERE          prodno = pur_rec.prodno AND
                        site_code = pur_rec.site_code AND
                  unit_treated = pur_rec.unit_treated AND
                  record_id_type = pur_rec.record_id_type;
      EXCEPTION
         WHEN VALUE_ERROR THEN
            DBMS_OUTPUT.PUT_LINE('Value error!');
            DBMS_OUTPUT.PUT_LINE('prodno = '||pur_rec.prodno||
               ', site_code = '||pur_rec.site_code||', unit_treated = '||pur_rec.unit_treated||
               ', record_id_type = '||pur_rec.record_id_type);

         WHEN NO_DATA_FOUND THEN
            numrecs := 0;

         WHEN TOO_MANY_ROWS THEN
            DBMS_OUTPUT.PUT_LINE('Too many rows error!');
            DBMS_OUTPUT.PUT_LINE('prodno = '||pur_rec.prodno||
               ', site_code = '||pur_rec.site_code||', unit_treated = '||pur_rec.unit_treated||
               ', record_id_type = '||pur_rec.record_id_type);

         WHEN OTHERS THEN
            DECLARE
               error_msg      VARCHAR2(300) := SQLERRM;
            BEGIN
               DBMS_OUTPUT.PUT_LINE('Other Error: ' || error_msg);
               DBMS_OUTPUT.PUT_LINE('prodno = '||pur_rec.prodno||
               ', site_code = '||pur_rec.site_code||', unit_treated = '||pur_rec.unit_treated||
               ', record_id_type = '||pur_rec.record_id_type);
            END;
      END;

      IF numrecs > 0 THEN
         lbs_per_unit := pur_rec.lbs_prd_used/pur_rec.acre_treated;

         ai_a_1000_200_flag := NULL;
         ai_a_2000_400_flag := NULL;
         prd_u_25m_flag := NULL;
         prd_u_50m_flag := NULL;
         prd_u_10md_flag := NULL;
         prd_u_50md_flag := NULL;
         nn1_flag := NULL;
         nn2_flag := NULL;
         nn3_flag := NULL;
         nn4_flag := NULL;
```

```
acre700_flag := NULL;

/**** CRITERION 1 ****/
/* If units treated are in acres, flag using criterion 1 */
IF pur_rec.unit_treated = 'A' AND ai_a_1000_200 > 0 THEN
   IF lbs_per_unit > ai_a_1000_200 THEN
         ai_a_1000_200_flag := 'Y';
   ELSE
      ai_a_1000_200_flag := 'N';
   END IF;
   IF lbs_per_unit > ai_a_2000_400 THEN
         ai_a_2000_400_flag := 'Y';
   ELSE
      ai_a_2000_400_flag := 'N';
   END IF;
END IF;

/* All other criteria only apply if there are more than 1 records
 * per set of records
 */
/**** CRITERION 2 ****/
IF numrecs > 1 AND prd_u_25m > 0 THEN
   IF lbs_per_unit > prd_u_25m THEN
         prd_u_25m_flag := 'Y';
   ELSE
      prd_u_25m_flag := 'N';
   END IF;
   IF lbs_per_unit > prd_u_50m THEN
         prd_u_50m_flag := 'Y';
   ELSE
      prd_u_50m_flag := 'N';
   END IF;
END IF;

/**** CRITERION 3 ****/
IF numrecs > 2 AND prd_u_10md > 0 THEN
   IF lbs_per_unit > prd_u_10md THEN
         prd_u_10md_flag := 'Y';
   ELSE
      prd_u_10md_flag := 'N';
   END IF;
   IF lbs_per_unit > prd_u_50md THEN
         prd_u_50md_flag := 'Y';
   ELSE
      prd_u_50md_flag := 'N';
   END IF;
END IF;

/**** CRITERION 4 ****/
IF numrecs > 1 AND nn1 > 0 THEN
   IF lbs_per_unit > nn1 THEN
         nn1_flag := 'Y';
   ELSE
      nn1_flag := 'N';
   END IF;
   IF lbs_per_unit > nn2 THEN
         nn2_flag := 'Y';
   ELSE
      nn2_flag := 'N';
   END IF;
   IF lbs_per_unit > nn3 THEN
         nn3_flag := 'Y';
   ELSE
      nn3_flag := 'N';
   END IF;
   IF lbs_per_unit > nn4 THEN
         nn4_flag := 'Y';
   ELSE
      nn4_flag := 'N';
   END IF;
```

```
         END IF;

         /**** CRITERION 5 ****/
         IF pur_rec.unit_treated = 'A' THEN
            IF pur_rec.acre_treated > 700 THEN
                  acre700_flag := 'Y';
            ELSE
               acre700_flag := 'N';
            END IF;
         END IF;

         INSERT INTO OUTLIER2002 VALUES
            (pur_rec.use_no, ai_a_1000_200_flag, ai_a_2000_400_flag,
             prd_u_25m_flag, prd_u_50m_flag, prd_u_10md_flag, prd_u_50md_flag,
             nn1_flag, nn2_flag, nn3_flag, nn4_flag, acre700_flag);
         COMMIT;
      END IF;
      END LOOP;
END;
/
show errors

CREATE UNIQUE INDEX OUTLIER2002_ndx ON OUTLIER2002
            (use_no)
      pctfree 5
      storage( initial 5M next 5M pctincrease 0)
      TABLESPACE NDX;
```